Bioinformatic analyses for T helper cell subtypes discrimination and gene regulatory network reconstruction

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Dr. rer. nat.
im Fach Informatik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
Humboldt-Universität zu Berlin

von
Dipl.-Inf. Stefan Kröger

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. habil. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter:
1. Prof. Dr. Ulf Leser
2. Prof. Dr. Joachim Selbig
3. Prof. Dr. Nils Blüthgen

eingereicht am: 28.02.2017
Tag der mündlichen Prüfung: 07.07.2017

Abstract

Within the last two decades high-throughput gene expression screening technologies have led to a rapid accumulation of experimental data. The amounts of information available have enabled researchers to contrast and combine multiple experiments by synthesis, one of such approaches is called meta-analysis. In this thesis, we build a large gene expression data set based on publicly available studies for further research on T cell subtype discrimination and the reconstruction of T cell specific gene regulatory events. T cells are immune cells which have the ability to differentiate into subtypes with distinct functions, initiating and contributing to a variety of immune processes. To date, an unsolved problem in understanding the immune system is how T cells obtain a specific subtype differentiation program, which relates to subtype-specific gene regulatory mechanisms.

We present an assembled expression data set which describes a specific T cell subset, regulatory T (Treg) cells, which can be further categorized into natural Treg (nTreg) and induced Treg (iTreg) cells. In our analysis we have addressed specific challenges in regulatory T cell research: (i) discriminating between different Treg cell subtypes for characterization and functional analysis, and (ii) reconstructing T cell subtype specific gene regulatory mechanisms which determine the differences in subtype-specific roles for the immune system. Our meta-analysis strategy combines more than one hundred microarray experiments. This data set is applied to a machine learning based strategy of extracting surface protein markers to enable Treg cell subtype discrimination. We identified a set of 41 genes which distinguish between nTregs and iTregs based on gene expression profile only. Evaluation of six of these genes confirmed their discriminative power which indicates that our approach is suitable to extract candidates for robust discrimination between experiment classes.

Next, we identify gene regulatory interactions using existing reconstruction algorithms aiming to extend the number of known gene-gene interactions for Treg cells. We applied eleven GRN reconstruction tools based on expression data only and compared their performance. Taken together, our results suggest that the available methods are not yet sufficient to extend the current knowledge by inferring so far unreported Treg specific interactions. Finally, we present an approach of integrating multiple data sets based on different high-throughput technologies to reconstruct a subtype-specific GRN. We constructed a Th2 cell specific gene regulatory network of 100 genes. While 89 of these are known to be related to Th2 cell differentiation, we were able to attribute 11 new candidate genes with a function in Th2 cell differentiation. We show that our approach to data integration does, in principle, allow for the reconstruction of a complex network.

Future availability of more and more consistent data may enable the use of the concept of GRN reconstruction to improve understanding causes and mechanisms of cellular differentiation in the immune system and beyond and, ultimately, their dysfunctions and diseases.

Zusammenfassung

Die Etablierung von Hochdurchsatz-Technologien zur Durchführung von Genexpressionsmessungen führte in den letzten 20 Jahren zu einer stetig wachsende Menge an verfügbaren Daten. Sie ermöglichen durch Kombination einzelner Experimente neue Vergleichsstudien zu kombinieren oder Experimente aus verschiedenen Studien zu großen Datensätzen zu vereinen. Dieses Vorgehen wird als Meta-Analyse bezeichnet und in dieser Arbeit verwendet, um einen großen Genexpressionsdatensatz aus öffentlich zugänglichen T-Zell Experimenten zu erstellen. T-Zellen sind Immunzellen, die eine Vielzahl von unterschiedlichen Funktionen des Immunsystems inititiieren und steuern. Sie können in verschiedene Subtypen mit unterschiedlichen Funktionen differenzieren.

Der mittels Meta-Analyse erstellte Datensatz beinhaltet nur Experimente zu einem T-Zell-Subtyp, den regulatorischen T-Zellen (Treg) bzw. der beiden Untergruppen, natürliche Treg (nTreg) und induzierte Treg (iTreg) Zellen. Eine bisher unbeantwortete Frage lautet, welche subtyp-spezifischen gen-regulatorische Mechanismen die T-Zell Differenzierung steuern. Dazu werden in dieser Arbeit zwei spezifische Herausforderungen der Treg Forschung behandelt: (i) die Identifikation von Zelloberflächenmarkern zur Unterscheidung und Charakterisierung der Subtypen, sowie (ii) die Rekonstruktion von Treg-Zell-spezifischen gen-regulatorischen Netzwerken (GRN), die die Differenzierungsmechanismen beschreiben. Die implementierte Meta-Analyse kombiniert mehr als 150 Microarray-Experimente aus über 30 Studien in einem Datensatz. Dieser wird benutzt, um mittels Machine Learning Zell-spezifische Oberflächenmarker an Hand ihres Expressionsprofils zu identifizieren. Mit der in dieser Arbeit entwickelten Methode wurden 41 Genen extrahiert, von denen sechs Oberflächenmarker sind. Zusätzliche Validierungsexperimente zeigten, dass diese sechs Gene die Experimenten beider T-Zell Subtypen sicher unterscheiden können.

Zur Rekonstruktion von GRNs vergleichen wir unter Verwendung des erstellten Datensatzes 11 verschiedene Algorithmen und evaluieren die Ergebnisse mit Informationen aus Interaktionsdatenbanken. Die Evaluierung zeigt, dass die derzeit verfügbaren Methoden nicht in der Lage sind den Wissensstand Treg-spezifischer, regulatorsicher Mechanismen zu erweitern. Abschließend präsentieren wir eine Datenintegrationstrategie zur Rekonstruktion von GRN am Beispiel von Th2 Zellen. Aus Hochdurchsatzexperimenten wird ein Th2-spezifisches GRN bestehend aus 100 Genen rekonstruiert. Während 89 dieser Gene im Kontext der Th2-Zelldifferenzierung bekannt sind, wurden 11 neue Kandidatengene ohne bisherige Assoziation zur Th2-Differenzierung ermittelt. Die Ergebnisse zeigen, dass Datenintegration prinzipiell die GRN Rekonstruktion ermöglicht.

Mit der Verfügbarkeit von mehr Daten mit besserer Qualität ist zu erwarten, dass Methoden zur Rekonstruktion maßgeblich zum besseren Verstehen der zellulären Differenzierung im Immunsystem und darüber hinaus beitragen können und so letztlich die Ursachenforschung von Dysfunktionen und Krankheiten des Immunsystems ermöglichen werden.

Acknowledgement

Finishing this PhD was a crucial milestone in my life – with periods of euphoria and despair I am very thankful and happy to have reached this goal.

My sincere gratitude goes to my supervisor Prof. Ulf Leser for his constant support, for his thought-provoking impulse and for encouraging me to overcome all scientific challenges. I gratefully thank Ria Baumgraß for offering me the entry position into the field of bioinformatics and immunology as well as for supporting me for so long. My sincere gratitude goes to Prof. Joachim Selbig and Prof. Nils Blüthgen for reviewing my thesis.

I am very thankful to my fantastic colleagues at the Humboldt-Universität zu Berlin and at the Deutsche Rheuma-Forschungszentrum Berlin for many helpful scientific discussions, and for their constant support and expertise. My special thanks goes to Samira Jaeger, Karin Zimmermann, Astrid Rheinländer and Philippe Thomas from the HU-Berlin and René Riedel, Melanie Venzke, Manja Jargosch and Martin Karl from the DRFZ. I am so proud to have worked with you all and it is now an honor for me to call you friends.

My indescribable thanks go to Katja Grabowski for her support, her help, her patience and her trust in me and in our life together. I could have never made it without her. And last but not least, I would like to sincerely thank my parents, my friends all the people that encouraged me to study, to start this PhD project and to keep going until today.

Contents

1. Introduction **1**
- 1.1. Goals and Contributions . . . 2
- 1.2. Outline of this Thesis . . . 3
- 1.3. Own prior work . . . 4

2. From gene expression to regulation in the immune system **7**
- 2.1. The mammalian immune system and immune cells . . . 7
- 2.2. T cells and T cell diversity . . . 8
  - 2.2.1. Fate determination and differentiation in T helper cells . . . 9
  - 2.2.2. T helper cell subtypes . . . 9
- 2.3. Gene regulation . . . 12
  - 2.3.1. Gene expression . . . 13
  - 2.3.2. Transcription factors and gene regulation . . . 14
- 2.4. Technologies and tools for sensing gene expression . . . 14
  - 2.4.1. Microarray technology . . . 14
  - 2.4.2. Next generation sequencing . . . 16

3. Meta-analysis of large gene expression experiments from public data repositories **21**
- 3.1. Motivation . . . 22
  - 3.1.1. Key issues for meta-analysis . . . 22
- 3.2. Data source for building meta expression experiments . . . 23
  - 3.2.1. Online data repository . . . 23
  - 3.2.2. Data curation . . . 25
- 3.3. Related work . . . 26
  - 3.3.1. Comparison of methods for meta-analysis of gene expression data . 26
  - 3.3.2. Types of meta-analysis . . . 27
  - 3.3.3. Existing meta-analysis approaches and implementations . . . 29
- 3.4. Assembling gene expression experiments for meta-analysis . . . 31
  - 3.4.1. Data collection . . . 31
  - 3.4.2. Retrieving meta information . . . 32
  - 3.4.3. Retrieving raw data . . . 33
  - 3.4.4. Meta expression set . . . 34
  - 3.4.5. Identifier mapping . . . 35
  - 3.4.6. Renormalization of assembled meta experiment . . . 36

3.5. Results . . . 39
3.5.1. Data set assembly and re-normalization . . . 39
3.5.2. Differential expression analysis . . . 39
3.6. Summary . . . 41

4. Ensemble feature selection for Treg cell subtype marker identification 45
4.1. Motivation . . . 45
4.2. Feature selection . . . 46
4.2.1. Methods for feature selection . . . 46
4.2.2. Feature selection for gene marker detection . . . 49
4.3. Feature set evaluation using linear SVM classifier . . . 50
4.3.1. Classification of imbalanced data sets . . . 51
4.4. Detecting robust markers by ensemble feature selection . . . 51
4.4.1. Data preparation & preprocessing . . . 53
4.4.2. Data partitioning and splitting into training and test sets . . . 53
4.4.3. Feature selection methods application and candidate merge . . . 53
4.4.4. Subset evaluation using SVM classifier . . . 56
4.4.5. Experimental validation of detected Treg subtype marker gene candidates . . . 57
4.4.6. Discussion . . . 60
4.5. Summary . . . 62

5. Reconstruction of gene regulatory networks 63
5.1. Motivation . . . 63
5.2. Methods for GRN reconstruction . . . 65
5.2.1. Formal definition of gene regulatory networks (GRNs) . . . 65
5.2.2. Co-expression or correlation networks . . . 66
5.2.3. Bayesian networks . . . 68
5.2.4. Information theory approaches . . . 71
5.2.5. Boolean networks . . . 72
5.2.6. Ordinary Differential Equations . . . 73
5.2.7. Adaptions of reconstruction algorithms . . . 74
5.2.8. Integration of prior knowledge to GRN reconstruction . . . 74
5.3. Tools implementing GRN reconstruction . . . 76
5.4. Overview of selected GRN reconstruction methods and tools . . . 79
5.5. Application of reconstruction tools to infer a Treg cell specific GRN . . . 81
5.5.1. Evaluation databases (silver standards) . . . 81
5.5.2. Tool testing strategy . . . 83
5.5.3. Results – Performance of GRN reconstruction tools on Treg cell expression data . . . 83
5.5.4. Discussion of the results from GRN reconstruction . . . 89
5.6. Summary . . . 91

6. Data integration for constructing T cell subtype specific regulatory networks 93
6.1. Motivation . . . 93
6.2. Integrating ChIP-seq, RNA-seq, Microarrays data . . . 94
6.2.1. Processing raw data from ChIP-seq, RNA-seq and microarray experiments . . . 94
6.2.2. Data integration procedure . . . 97
6.3. Reconstruction of a *Stat6*-centered network for Th2 cells . . . 98
6.3.1. Preparation and analysis of publicly available data to reconstruct a gene regulatory network . . . 98
6.3.2. Th2 cell specific gene regulatory network assembly . . . 101
6.3.3. Implications from the reconstructed network . . . 102
6.3.4. Evaluation of the network . . . 102
6.3.5. Discussion of the presented integration strategy . . . 104
6.4. Summary . . . 104

7. Summary & Future work 107
7.1. Summary . . . 107
7.2. Future directions . . . 109
7.2.1. Meta-analysis . . . 109
7.2.2. Cell specific surface marker identification . . . 109
7.2.3. Gene regulatory network reconstruction . . . 110

A. Appendix 113
A.1. References to microarray experiments integrated into meta-analysis . . . 113

1. Introduction

Since the beginning of modern natural sciences and especially biology, strategies and methods to uncover natural phenomena were mainly based on observation and interpretation. Technological inventions like microscopy and x-rays, mass spectrometry and DNA sequencing opened new and magnitudes deeper views inside objects and organisms. Later simulation and hypotheses testing via models became popular. In the last century, the availability of computing power and computer-based, so called *in silico* experiments enabled the construction of complex virtual models. The ongoing development of technologies and sensing devices increased the amount of experimental data and in consequence the need for appropriate computational analysis methods.

In biology and related life science fields, computational analysis became relevant in the second half of the 20th century with the genomic era. Its breakthrough came with the advent of the so called post-genomic era and the establishment of DNA sequencing and other high-throughput omics-technologies leading to the rapid accumulation of experimental data [126, 144]. In contrast to conventional experimental "wet-lab" techniques, more fine grained and deeper insights were achievable with less effort. In this time, the focus of bioinformatic tools changed from experimental observations and data collection to results analysis. New methods for comparative analysis of genomic data became indispensable.

Baldi et al. described in [18] two phases of computational support in biology, termed computational biology or bioinformatics. The first phase of computational biology was focused on data analysis like sequencing data. The second phase relates to sophisticated integration of highly diverse kinds of data. Data integration based on a variety of newly accessible experimental techniques, many of which are capable of data generation at different levels, i.e. cell, tissue, organ, organism, or even population level. Until today, the amount of accumulated biological data has increased well beyond petabyte [218]. Ever since, large amounts of data became publicly available, researchers aimed to contrast and combine multiple experiments by *synthesis* so called meta-analysis [48]. Since all experiments are unique in their specific settings and conditions comparing different experiments and results is difficult. Meta-analysis has evolved into an essential tool to discover candidate biomarkers, understand biological processes or resolve conflicting conclusions from different experiments [48, 146]. Meta-analysis rapidly became an essential tool of exploratory analysis, e.g. during drug response analysis or disease gene identification [397]. In addition, meta-analysis enables the aggregation of related experiments to large-scale data sets, required for gene regulatory network (GRN) inference [236], that aims to uncover the regulatory gene-gene interactions and regulatory mechanisms between genes.

While the concept of genes and gene regulation was discovered more than 70 years ago

[84], many questions regarding the specific interactions, so called gene regulations are still unanswered. Multiple such regulations can be presented as interactions networks or so called GRNs.

Knowledge about the GRN is crucial to understand the pathogenesis of Mendelian disorders or systematic dysfunctions of organisms, which can be caused by alterations and disruptions of the underlying regulatory mechanisms [14, 264]. For instance, some malfunctions of the immune system can be tracked down to transcription factor (TF)–initiated and cytokine–initiated gene regulatory events [117]. Alterations of the gene regulatory program can cause pathological cell development and changes in resulting autoimmunity with consequences like rheumatoid arthritis or allergies. During the last 30 years, researchers aim to uncover the regulatory mechanism of the immune system, for example the regulation of T helper cell differentiation [75]. While, it is known, that T helper cells differentiate into different subtypes that maintain specific function to the immune system the regulatory mechanism behind are not fully understood and even unknown for certain subtypes.

In this thesis we aim to contribute to the problem using machine learning algorithms applied to a large gene expression data set. In particular we study two challenges, that arose in these years. First, we aim to develop a method that allows for the identification of discriminating surface marker for different regulatory T helper cell subtypes. We implement a machine learning approach to discriminate between induced and natural regulatory T helper cells. Second, we aim to compare methods for reconstructing T helper cell specific GRNs from gene expression data to elucidate important gene regulatory interactions during regulatory T helper (Treg) cell differentiation by presenting two general approaches. In Chapter 5, we compare existing algorithms for GRN reconstruction to extend the number of known Treg cell specific gene regulatory interactions. In Chapter 6 we present a data integration approach to reconstruct a GRN using results and data from existing experiments, public databases and literature. We present a sophisticated data integration strategy to recover and infer gene regulatory interactions between the master transcription factor Stat6 and other transcription factors and cytokines during Th2 cell differentiation.

1.1. Goals and Contributions

This work is embedded in the context of understanding the role of T helper cell for the mammalian immune system. We show that the integration of publicly available data combined with the application of GRN reconstruction algorithms and machine learning enables a more detailed understanding of the molecular mechanisms in T helper cells and provides new hypotheses about specific gene regulatory interactions.

The four major topics of this thesis are (i) the application of meta-analysis strategies on publicly available gene expression data for (ii) the identification of marker genes, (iii) the inference of gene regulatory networks from gene expression data and (iv) the reconstruction of gene interactions by integrating experimental data and established knowledge from different sources.

The specific contribution of this work are:

- Collect and assemble a set of publicly available T helper cell subtype specific microarray gene expression experiments. We present different approaches for meta-analysis and present our own, which is based on the integration of raw data instead of published results. We briefly describe problems on integrating heterogeneous data sets and explain how to reduce data set specific bias. Next, we compare this strategy to "t-test"-based differential expression analysis, and use gene set annotations to show the meta-analysis' capability of reducing the differences between experiments coming from different studies, while preserving the T helper cell specific expression characteristics of the genes.

- We develop a pipeline for applying feature selection methods to gene expression profiles to extract candidate genes that code for surface marker proteins. Thereby we aim to enable the experimental distinction of different Treg cell subtypes and thus allow for deeper biological function analysis. As a consequence thereof, we implement an *in silico* evaluation strategy using Support-Vector-Machines.

- We analyze and compare available methods for GRN reconstruction from gene expression data. We run a selection of tools without providing additional prior knowledge to the reconstruction methods. In contrast to other studies, we apply the methods to noisy, murine expression data. We show that the performance of tools without additional knowledge and with heterogeneous data quality is rather poor.

- We demonstrate a gene regulatory network around Stat6 – the master transcription factor of Th2 cells. Thereby we demonstrate the power of integration strategies using a variety of RNA- and DNA-sequencing and microarray data, as well as literature information. Next to a number of known interactions of important transcription factors and cytokines during Th2 cell differentiation, we extract 11 targets of Stat6 with so far unreported functions in Th cell differentiation.

1.2. Outline of this Thesis

Chapter 2 provides the biological background relevant throughout this work. It outlines the core structure of the mammalian immune system, and explains the principles of T helper cell function - the biological context of this thesis. Subsequently, we introduce the concept of gene expression and gene regulation. We conclude this chapter by introducing high-throughput detection technologies, like microarray technology and Next Generation Sequencing for RNA and DNA.

Chapter 3 presents our approach to the aggregation of many publicly available Treg cell specific gene expression experiments to form a large meta-experiment for further application. The goal is to compensate the lack of large homogenous data sets during the application of subsequent data analysis strategies, like machine learning

for marker detection (Chapter 4) or gene regulatory network reconstruction based on expression data (Chapter 5). Finally, we perform and evaluate the meta-analysis of selected Treg cell specific microarray experiments.

Chapter 4 describes a machine learning approach to extract Treg cell specific surface markers using the assembled meta-expression set. We give an overview about feature selection methods and their application for gene marker detection, followed by describing Support Vector Machines (SVMs) a supervised machine learning algorithm for classification. The SVM classifier is used to test marker qualities of extracted candidate genes.

Chapter 5 presents the methods for GRN reconstruction. At first, we explain the importance of GRNs to unravel the complex topology of gene-gene interactions with T-helper cells. Second, we give a broad overview about existing algorithmic approaches and tools for GRN reconstruction on whole-genome scale from gene expression data. Third, we present the application of 11 of these tools to the meta-expression set. We evaluate the inferred networks using public databases and compare their respective performance.

Chapter 6 describes an alternative approach for reconstructing gene interaction networks by data integration for a specific T-helper cell subtype on small scale. Using the example of Th2 cells we construct a network based on time-series gene expression data from sequencing and microarrays, TF-binding data from high-throughput DNA sequencing experiments as well as public database information. The reconstructed network is specific to Stat6 - the master transcription factor of Th2 cells.

Chapter 7 summarizes this thesis and the main contributions. Finally, it gives an outlook to future work.

1.3. Own prior work

Some parts of this thesis are based on work which has been published previously in peer-reviewed publications.

Chapter 3 and 4 are based on a manuscript, in preparation [198]. Parts of the article are mentioned in this thesis, are the specific experiment selection in Chapter 3 and the wet-lab experiments in Chapter 4, which all were performed by M. Ventzke and M. Karl from the German Rheumatism Research Center in Berlin. Other parts, like the meta-analysis, the concept, design and implementation of the feature selection and *in silico* evaluation can be attributed to the author of this thesis. The resulting manuscript was drafted by all authors. U. Leser additionally supervised the whole project, while R. Baumgrass supervised the biological experiments.

Chapter 6 presents the contributions of the author to the publication [168]. M. Jargosch and S. Kröger are equally contributing first authors in the referenced literature. The author contributed the concept and design for the data integration and performed parts of the data analysis (microarray, RNA-seq, ChIP-seq), partly in collaboration with

E. Gralinska and U. Klotz. Experiments in this study were selected, conceived and designed by M. Jargosch, R. Baumgrass, Z. Fang and W. Chen. R. Baumgrass, U. Leser, J. Selbig and D. Groth supervised the project. M. Jargosch, R. Baumgrass and S. Kröger drafted the manuscript.

2. From gene expression to regulation in the immune system

The following chapter presents the biological and technical background regarding methods and contexts of the subsequent chapters.

First, we describe the immune system and its role in mammalian organisms, followed by a brief introduction to T cells and specific subtypes as well as the process of T cell differentiation. We mainly focus on two subtypes: Treg and T helper 2 (Th2) cells, as these are addressed in Chapters 3, 4 and 5 (Treg cells) and Chapter 6 (Th2 cells). Later we introduce the phenomena of gene expression and gene regulation and describe transcription factors and their differential roles in T cells development and fate determination. The fourth section of this chapter elaborates the technical bases of determining gene expression and gene regulation by microarray and state of the art sequencing technology.

2.1. The mammalian immune system and immune cells

The general task of immunity as conveyed by the immune system of higher vertebrates is to combine a set of mechanisms for discrimination between self[1] and non-self for defense against pathogens. The immune system's major function is to prevent the organism from intrusion of foreign organisms whilst at the same time controlling its own defensive response in order to prevent damage to the organism itself.

Higher vertebrates like mammals employ complex immune systems comprised of a large variety of cells from different origins with differential functions and a large array of different mechanisms. The immune system can be divided into two general concepts, the innate and adaptive immune system. Both employ different strategies for the recognition of pathogen-related patterns (for details about the two concepts, see e.g. [1, 167, 269]). Cells, like macrophages, granulocytes or natural killer cells make up the innate immune system. These cells utilize receptors which recognize evolutionary conserved molecular patterns associated with pathogens. Such receptors, e.g. the Toll-like receptors (TLRs), the best known and well-studied examples, are genetically encoded and can be employed by any cell which expresses the gene [211].

B cells and T cells are the cells of the adaptive immune system and can employ a wide range of molecular mechanisms against the pathogens, i.e., resulting in production of mediators of inflammation like cytokines. They employ antigen-receptors generated from randomly selected and combined gene segments. Antigen binding of a receptor activates

[1] e.g. belonging to and produced by the organism

B and T cells and leads to the proliferation of antigen-reactive cells. Efficient immune responses require the proliferation of initially very few antigen-reactive cells, so called clonal expansion and selection of cells with enhanced antigen-binding capacity. The antigen-directed (therefore: adaptive) immune responses are much slower than innate responses (weeks rather than minutes or hours). After driving an adaptive immune response the pathogens have been cleared from the system, some antigen-responsive cells are retained as memory cells. The presence of these cells allows for a much faster response of the adaptive immune system upon a second confrontation with a known antigen. This property of the adaptive immune system is referred to as immunological memory and is put to use in vaccination.

Adaptive Immune System

Lymphocytes, the cells of the adaptive immune system, originate from a common progenitor cell in the bone marrow and mature in the bone marrow (B cells) or in the thymus (T cells), respectively. T cells are divided into cytotoxic and T helper (Th) cells: Cytotoxic T cells mediate killing of host cells which are infected with intracellular pathogens whereas Th cells are required for initiation and control of cellular actions of the immune system, e.g. the B cell response leading to the production of antigen or cytotoxic T cell action.

Immune Tolerance

One key property of the immune system is the immune tolerance. The term refers to the immune system's ability to differentiate between host (so-called: self) antigens originating from the organism itself, and foreign antigens in such a way that potent immune responses are only targeting non-self antigens.

The antigen receptors of the adaptive immune system are generated randomly. Therefore they can, in principle, not only bind foreign but also self-antigens. This situation is called auto-immunity and is potentially deleterious to an organism. Auto-immunity can give rise to auto-immune diseases, such as rheumatoid arthritis, diabetes, systemic lupus erythematosus with varying levels of strength. Auto-immunity is tightly controlled by a variety of mechanisms such as the deletion of immune cells reactive to self-antigen during development and maturation in the bone marrow (B cells) or T cells in the thymus (central tolerance). To control and confine auto-immunity Th cells can differentiate into Treg cells in the secondary lymphoid organs (e.g. spleen, lymph nodes, or tonsils) or in the periphery. Peripheral Treg cells have the potential to inhibit immune responses to self-antigens by various mechanisms (peripheral tolerance).

2.2. T cells and T cell diversity

T cell development and selection for central tolerance take place in the Thymus where T cells differentiate into either cytotoxic T cells, generally identified by the expression of cluster of differentiation 8 (CD8) [382] or into Th cells generally expressing cluster of

differentiation 4 (CD4). Th cells provide signals which allow for full activation of innate immune cells, cytotoxic T cells or B cells [344]. CD4 expressing T cells can further differentiate into several Th cell subsets or into regulatory T cells (for a simplified scheme see Fig. 2.1).

Th cells express the CD4 co-receptor and have a major role in controlling and regulating the immune system by providing necessary co-stimulatory signals for the action of other white blood cells [167]. The activation of T cells is mainly driven via T cell receptor (TCR) stimulus and requires co-stimulation. T cell activation takes place in lymphatic tissues such as as the spleen or lymph nodes of the periphery and after contact with Antigen-presenting cells (APCs). To prevent inappropriate immune response to self-antigens (autoimmunity), T cells become anergic and later activation is unlikely. After activation, T cells change the composition of surface molecules, i.e. molecules which are involved in their respective functions. This also enables identification of the T cell subtypes by differences in molecules present on their cell surface (see Chapter 4).

2.2.1. Fate determination and differentiation in T helper cells

The ability of proliferation into different subtypes and functional plasticity of naïve T (Th0) cells is a prerequisite for T cell immunity. Then acquisition of lineage-specific T cell effector functions is linked to a proliferative response, suggesting that T cell activation drives differentiation programs which facilitate effector gene expression [300]. In recent years, much effort has been spent to investigate the plasticity resp. the ability of re-differentiation of T cells from one to another subtype (for details on plasticity see e.g. [111, 206, 268, 378, 405]). The differentiation of naïve CD4+ T cells (Th0) into Th cell subtypes is regulated by specific cytokine milieus and complex TF networks [168]. After activation, Th0 cells are able to differentiate into one of six distinct T cell subsets, defined by the presence of cytokines and subtype specific TFs [268]. Other subtype definitions, like Th22, Th9 or Th40 cells [233] are not of interest in this thesis.

2.2.2. T helper cell subtypes

Here, we focus on four of these six subtypes, namely Th1, Th2, Th17 and Treg cells and the process of T cell fate determination. Furthermore, we briefly describe the role of each subtype during immune response.

Type 1 T helper (Th1) cells

Th1 cells develop under the presence of interleukin 12 (IL-12) and Interferon γ (IFNγ), both are the critical cytokines initiating the downstream signaling cascade to develop Th1 cells. IL-12 is secreted by APCs upon activation. IFNγ is produced by natural killer cells (NKCs) which in turn are induced by IL-12. The master TF of Th1 cells is the T-box TF (T-bet). T-bet drives the production of IFNγ, and suppresses Th2 and Th17 programs. Th1 cells host immunity effectors against intracellular bacteria and protozoa by mediating immune responses by induction of proliferation of cytotoxic $CD8^+$ T cells and macrophages.

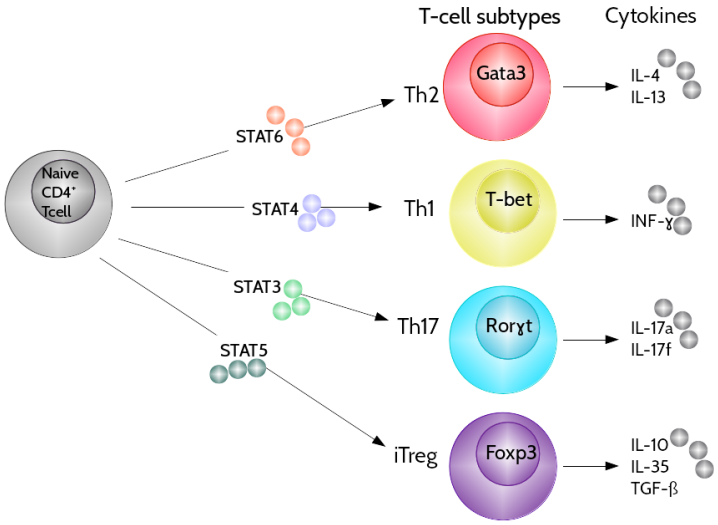


Figure 2.1.: **Development of naïve T (Th0) cells into T helper cell lineages/ subtypes.** Process of T cell differentiation is mainly driven by presence and absence of lineage specific cytokines (written on the arcs) and so called master TFs (written in the cells). Cytokines at the right side are those, expressed by the activated cells.

Type 2 T helper (Th2) cells

Naïve T helper cells can differentiate into Th2 cells in the presence of Interleukin-4 (IL-4) [91, 167]. These cells produce various cytokines (IL-4, IL-5, IL-6, IL-9, IL-13, and IL-17E (IL-25)) to mediate their functions, which encompasses mainly the activation and maintenance of the humoral, or antibody-mediated, immune response against extracellular parasites, bacteria, and toxins. IL-4 itself drives GATA3 expression through the TF Stat6. GATA3 is the master TF of Th2 cell differentiation (see Fig 2.1). Subsets of immune cells are stimulated and recruited by Th2 cells, such as eosinophil and basophil granulocytes. The activation of Th2 cells is a major factor for the exacerbation of Type-1 allergies (immediate hypersensitivity reactions), autoimmune reactions such as chronic graft-versus-host disease, progressive systemic sclerosis, and systemic lupus erythematosus [91].

Regulatory T (Treg) cells

The ability of the immune system to balance and regulate immune responses is largely driven by $Foxp3^+$ regulatory T (Treg) cells as part of the immunological tolerance (see Fig 2.2). Therefore, Treg cells are not only important objects of study but also targets for manipulation, as reduced amounts or dysfunction of these cells can lead to autoimmune-diseases. Over-representation or hyperfunction of Treg cells can lead to cancer [265, 368].

As early as 1995, Sakaguchi et al. showed the importance of $CD4^+CD25^+$ Treg cells for the prevention of autoimmune diseases [304]. A particular important feature of

Foxp3$^+$ regulatory T cells is the expression of the master TF Foxp3 [149]. In human, functional mutations in this TF can result in a loss of Treg cells and development of the immunodysregulation polyendocrinopathy enteropathy X-linked syndrome (IPEX, first described in [280]).

Foxp3$^+$Treg cells show a certain degree of diversity and different subpopulations have been identified among the pool of *in vivo* occurring Foxp3$^+$ Treg cells, also designated as natural Treg (nTreg) cells [49, 101, 158]. Discriminated by their origin, two important subtypes are thymus derived Treg (tTreg) and peripherally induced Treg (pTreg) cells, reviewed in [312, 324]. The differentiation of tTreg cells occurs after positive selection in response to high affinity TCR signals to self-antigens and is dependent on γc receptor signaling and to some extent also TGF-β [59]. tTreg cells make up a substantial fraction of the nTreg pool [151, 172]. In contrast, pTreg cells differentiate from conventional CD4$^+$ T cells in response to low-doses of antigen in an anti-inflammatory environment and dependent on TGF-β, which occurs especially at sites of antigen encounter, i.e., mucosal tissues like the gastrointestinal tract and the lungs. Regarding functional differences, tTreg cells are critical for the prevention of autoimmunity whereas pTreg cells are induced to narrow more localized inflammatory responses, esp. against foreign antigens [76].

Furthermore, naïve CD4$^+$ T cells differentiate *in vitro* into Foxp3$^+$ iTreg cells after activation in the presence of TGF-β [60]. However, these so-called iTreg cells do not show the same features as their natural counterparts, especially in terms of lineage stability and function. While nTreg cells are in general more efficient in their suppressive function, iTreg cells are highly efficient in controlling the onset of autoimmunity in a mouse model of autoimmune gastritis [389]. Additionally, Huang et al. showed that iTregs are more tolerogenic in an asthma model [156].

For years, Treg cells have been important subjects of medical research due to their crucial role for the adequate regulation of the immune system. Such research would benefit greatly from reliable surface markers allowing the effective discrimination of iTreg and nTreg cells. Being able to perform such discrimination would open the door for detailed functional studies of both cell subtypes and their respective roles within the immune system. Results from such studies could have immediate clinical implications [265, 327]. Also, as described in Chapter 4 a clear definition of nTreg and iTreg cells is required for cell transfer therapies [256, 350].

Type 17 T helper (Th17) cells

Naïve T helper cells differentiate under the presence of TGF-β and Interleukin 6 (IL-6) and Interleukin 23 (IL-23) to T helper 17 (Th17) cell [352]. Differentiated cells express Interleukin 17 (IL-17) which led to their name. The presence of TGF-β during differentiation relates these cells directly to regulatory T (Treg) cells as the IL-6 inhibits the differentiation of Treg cells. Th17 cells have been connected to autoimmune and inflammatory reactions as they react to pathogens and activate neutrophil granulocytes, which are among the early responding cells in the mammalian immune system [195]. Dysregulation of Th17 cells results in autoimmune disorder and inflammation. The

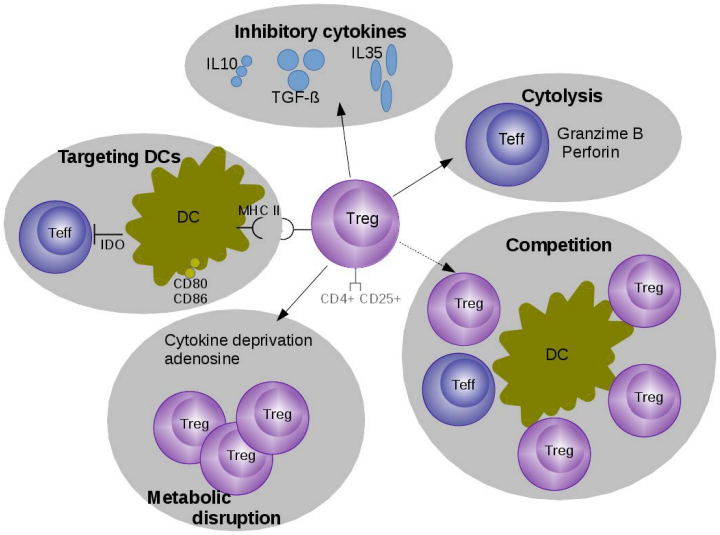


Figure 2.2.: **Putative mechanisms used by Treg cells.** Treg cells have multipe putative targets as described by Caridade et al. [51]. These include (1) targeting dendritic cells (DCs) – leading to weak or abrogated signals to naïve/effector T cells; (2) Metabolic disruption; (3) Competition – for critical cytokines, such as IL-2, or direct disruption of effector cell engagement with APCs; (4) Cytolysis – direct cytotoxic effect and consequent apoptosis of effector T cells or APCs; (5) Production of inhibitory cytokines – including IL-10, IL-35, and TGF-ß. Figure adapted from [51].

master TF of Th17 cells is STAT-3, which is activated by both IL-23 and IL-6, and, for example, binds to the IL-17 locus.

2.3. Gene regulation

Genes and the interaction of genes via their products are keys to the functions and mechanisms of organisms [7]. Gene regulation refers to processes controlling the expression of a gene. Gene regulation comprises direct mechanisms like gene-gene regulation and "implicit" mechanism like the circadian clock or the activity of so called enhancer elements [99, 208, 362]. By current knowledge not all regulatory effects can be fully explained, as it is assumed that further, so far unknown, factors are involved. Examples of such known factors influencing gene regulation during T cell differentiation are the presence of certain cytokines or the specific cytokine repertoire (physical amount), TF presents or others like Calcium influx during Th1/Th2 cell activation. Methods of gene expression tracking and comparative gene expression analysis using microarray or sequencing technology may help to elucidate regulatory mechanisms at transcriptome level. Thereby it is of importance that such analyses are performed under comparable physical states and conditions [17, 223].

2.3.1. Gene expression

Gene expression is the process of transcribing deoxyribonucleic acid (DNA) coding a gene to ribonucleic acid (RNA), before translation into a functional gene product (see Fig 2.3). Gene products are proteins or functional RNAs, respectively non-protein coding genes like transfer RNA (tRNA) or small nuclear RNAs (snRNAs) [7].

As described in Fig 2.3 gene expression consists of mainly four steps. First, DNA is transcribed into RNA. Second, the RNA is spliced to messenger RNA (mRNA). During splicing certain coding regions (exons) are kept and non-coding region (introns) are removed. The remaining exons are spliced together. Third, the spliced elements are exported into the cytoplasm after preparation . In the last step, the mRNA is translated into an amino acid sequence making up the protein.

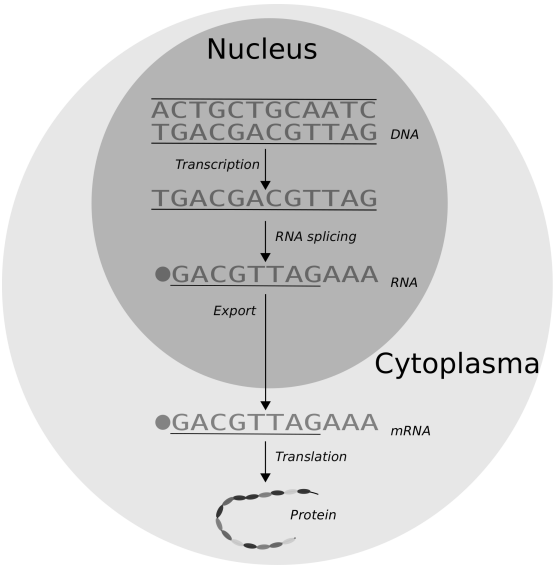


Figure 2.3.: **Gene expression of eukaryotic organ is initiated in the cell nucleus.** DNA is transcribed into RNA. Next RNA is spliced and transported from cell nucleus to the cytoplasm and finally translated to a protein coding amino acid sequence.

Notably, manipulation and error can act at every step of gene expression, i.e., transcriptional control during transcription of DNA to RNA, control of RNA processing in splicing (alternative splicing, siRNA), during the transport from the nucleus to the cytoplasm (tRNA control), via mRNA degradation (miRNA), or translational control during or before translation into protein [7]. Regulatory effects result from the complex interplay of multiple mechanisms, not exclusively controlled by TF or miRNA binding. Nevertheless, in this work we focus on TF as origin of regulatory events as TFs are still assumed to be one of the most impacting factors during gene regulation.

2.3.2. Transcription factors and gene regulation

TFs are proteins which regulate the transcription of other genes by binding to a certain DNA sequence and, thus, activating or repressing the expression of genes in the specific or flanking genomic regions. TFs can act together with other TFs to block or promote RNA polymerase , so called co-factorial binding or co-regulation via TF complexes (see Fig 2.4). In general, TFs are able to turn on or off the expression of other genes by controlling molecules which enable or disable transcription. Vaquerizas et al. estimated in 2009 about 1,700–1,900 TF-coding genes in the human genome [366], in mice there are about 1,700 genes coding for TFs [370]. Additionally, histone modification can influence binding and regulation by TFs as it allows or inhibits the binding of a TF at a specific binding site [140].

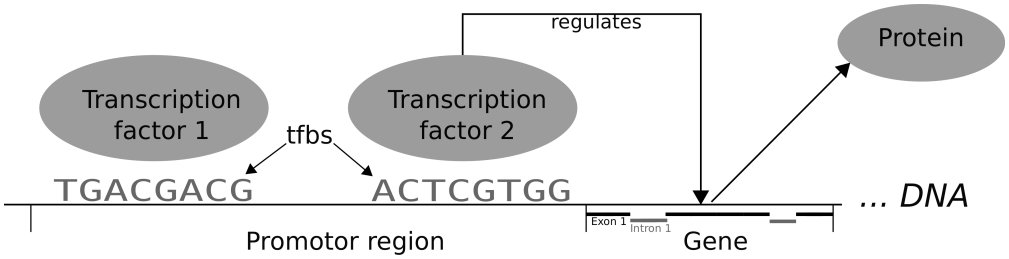


Figure 2.4.: **TF-binding in promoter region initiates gene transcription.** Binding of only one TF may cause a regulatory event, while binding of two can cause a different regulatory events [122]. The figure illustrates the phenomenon that gene regulation is cause by cooperatively acting of TFs, so called co-binding.

Transcription factors impact on T cell fate Various TF like TGF-β , TGIF and cytokines like Interleukin 4 (IL-4), Interleukin 6 (IL6) play a crucial role during Th cell fate decision as visualized in Fig 2.1. Their presence is necessary to initiate or inhibit the differentiation into Th cell subtypes. Therefore it is of high interest to uncover regulator mechanisms which lead to different expression levels of these factors.

2.4. Technologies and tools for sensing gene expression

2.4.1. Microarray technology

Microarrays allow the detection of several thousands of expressed messenger RNAs (mRNAs) molecules at the same time [310]. Defined genetic sequences (probes) are placed on small glass or plastic plates (wafers) to detect the expression of complementary sequences. Each probe is located at a specific location (spot) on these plates and detects a specific sequence. Probe sequences are either artificially created and standardized for a specific microarray type or individually prepared for so-called custom arrays. These probes are oligonucleotides (oligos), that are usually 13 to 25-mers (25 bases long

sequences), that are directly synthesized onto the wafer and refer to a gene or transcript. Multiple probes represent a single transcript and each probe itself consists of a Perfect-Match (PM) and a MissMatch (MM) oligo. While the PM has the exact same sequence as the target transcript, the MM is different by a single base substitution in the middle of the sequence. MMs enable better background and unspecific hybridization detection by comparing with the corresponding PMs. During the actual experiments, fluorescent antibodies are coupled with the sample sequences for optical recognition (scanning) of the amount of bound sequences for each spot. The measured fluorescence represents the amount (expression) of the bound sequences. The technology was developed in the 1980s, but real market values was reached first from the mid of the 1990s until today.

For this work the most important types of microarrays are single and two-color-spotted oligonucleotide microarrays. The standard protocol of performing a microarray expression experiment is shown in Fig 2.5.

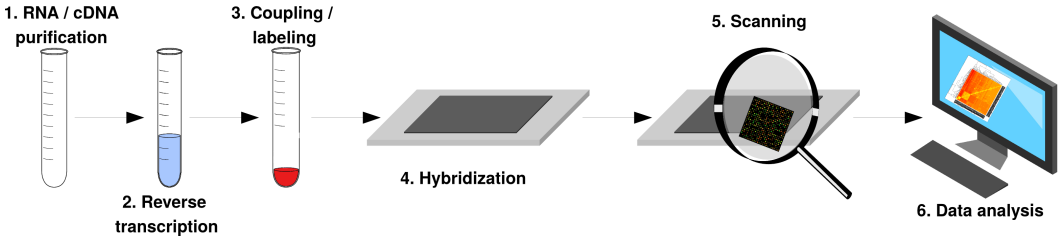


Figure 2.5.: **Schema of a microarray experiment from sample purification to data analysis.** The figure illustrates processing steps and their sequential order during a microarray experiments. (Necessary sample preparation steps before purification are not shown here.)

1. **RNA/ cDNA purification** includes extraction of the target cells and the extraction of RNA (purification). This step is important to avoid false positive signals through contamination (e.g. by foreign RNA fragments).

2. **Reverse transcription** is the process of generating complementary DNA (cDNA) from RNA using the enzyme reverse transcriptase (RT). These cDNA sequences bound to the specific probes during hybridization (step 4).

3. **Coupling** is the process of labeling the fragments, typically by fluorescent or radioactive labels. Labeling can be performed directly via specific fluorescent antibodies or indirectly via antibodies which in turn bind fluorescent antibodies. For two-color arrays labeling is performed before hybridization, while for single-color arrays the labeling takes place after hybridization.

4. **Hybridization and washing** is the actual process, where the cDNA fragments bind to the oligonuceotide probes on the microarray (hybridization). After hybridization residual unbound sequences and DNA fragments are removed (washing).

5. **Scanning** is the process of detection of emitted light or radiation. High-resolution pictures of the microarray show the level of gene transcription (gene expression) by the intensity of the specific light emitted from prior coupled fluorescent antibodies. The amount of light or radiation emitted is used to calculated raw expression values. Manufacturers often provide software, which (i) performs microarray type specific noise reduction and intensity normalization and (ii) translate the signals into numeric values.

6. **Data analysis** typically starts with background correction and normalization, which is mandatory for comparing expression values from different microarrays, like different samples, replicates, etc. Various proprietary and free protocols have been developed, Figure 2.6 illustrates an example analysis workflow. Until today, Microarray Analysis Suite 5 (MAS5) and Robust Multi-array Average (RMA) are two of the most widely used methods [132]. MAS5 performs normalization for each microarray independently using MM and PM. RMA performs normalization over multiple arrays based on PM only. RMA calculates expressions with corresponding P-values indicating the significance of erroneous [34], where MAS5 also provides flags for absence, minor presence and presence of a ProbeSet for a given transcript. Other examples of widely used analysis strategies are Gene Set Enrichment Analysis (GSEA) [343], cluster analysis to extract genome-wide expression patterns [92, 170, 342], or the application of gene network inference tools as described in Chapter 5.

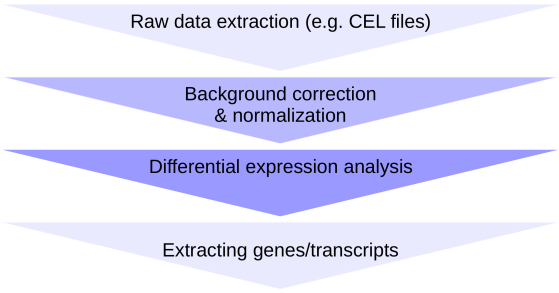


Figure 2.6.: **Microarray expression analysis workflow.** Various workflow variants similar to the presented (implementing different algorithms) exist.

2.4.2. Next generation sequencing

The first full organism, bacterial phage $\Phi X174$ was sequenced in 1977 by Sanger et al. [308] it consisted of just 5386 nucleotides encoding 11 genes. The human genome consists of 3,234.83 nucleotides on 23 chromosome pairs and encodes about 19,000 protein coding genes [96], which illustrates that high-throughput methods were necessary to sequence the entire human genome. The task of sequencing is to determine the exact order of the four bases (adenine, guanine, cytosine, and thymine) in a DNA or RNA

strand. In the beginning, two-dimensional chromatography was used to determine single bases in extracted strands, while modern sequencing devices enable automated determination of the full sequence from a small RNA/DNA sample. The era of high-throughput sequencing started in the mid 1990s, and the first commercial sequencer (GS20) was available in 2005[367]. At this time technology was expensive and error-prone. Over the years, sequencing technology and necessary computation power became cheaper by magnitudes (see Fig 2.7). Some of the most common Next Generation Sequencing (NGS) technologies are Illumina, 454 pyrosequencing, Solexa, SOLiD, Polonator and HeliScope sequencing [321]. NGS allows for detecting the dynamics of genome-wide expression without specific assumptions, like viral genes [50, 287, 313].

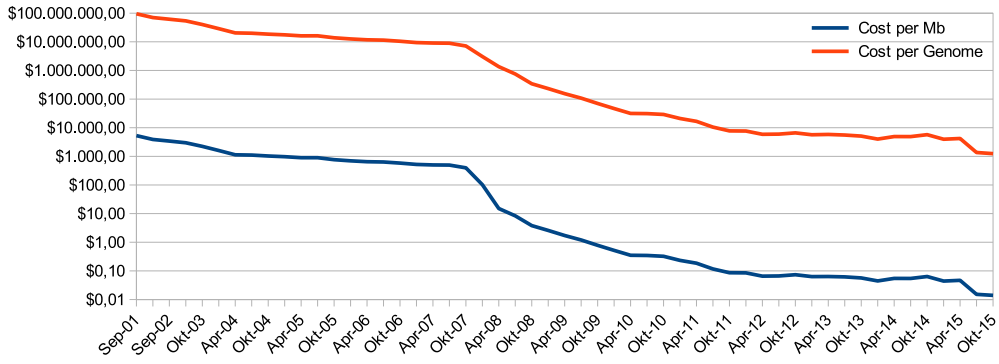


Figure 2.7.: **The decrease of sequencing costs per human genome over time.** The orange line additionally shows the cost per raw Megabases of DNA sequences. Data taken from the NHGRI Genome Sequencing Program (GSP) [197].

Chromatin immunoprecipitation DNA-sequencing (ChIPseq)

Beside sequencing of complete organisms, DNA sequencing (DNA-seq) can also be used to analyze specific changes during different states or under various conditions. As research draw attention to the function of specific regions and DNA sequence in the genomes, the interest in targeting such was initiated. DNA-seq can shed light on many DNA alterations and binding effects when combined with Chromatin Immunoprecipitation (ChIP) to measure protein binding to specific DNA regions, called Chromatin Immunoprecipitation sequencing (ChIP-seq). The detection if a specific protein (TF) can bind to a gene (TF target) or its surrounding region on the DNA of protein is the major application of ChIP-seq. Thereby, the TF binding regions are isolated as DNA fragments (ChIP) and sequenced (seq). Subsequent analysis reveals if the target gene (or surrounding genomic regions) contain such sequences (binding motifs). The sequence isolation is performed using specific antibodies that bind the TF or specific genomic region of interest (see Fig 2.8).

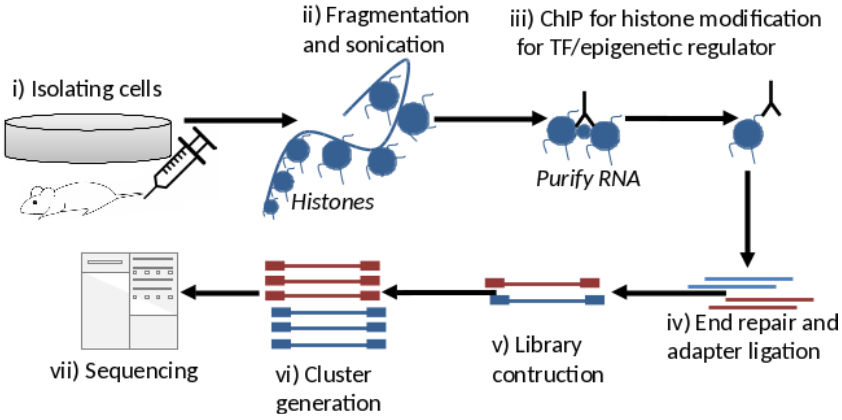


Figure 2.8.: **Illustration of the processing workflow for ChIP-seq experiments.** ChIP-seq procedure consists of i) the isolation of the target samples (here immune cells), ii) fragmentation of the isolated DNA, followed by iii) the chromatin immunoprecipitation with the subsequent RNA purification, iv) adapter ligation and end repair of the isolated sequences, v) library preparation and vi) clustering the targeted RNAs following by vii) the actual sequencing and subsequent analysis of detected reads. The figure is adapted from Kidder et al. [183].

After performing the chromatin-immunoprecipiation and subsequent sequencing of extracted DNA fragments the sequencing machine generates files that contain the raw reads. These raw reads are used for the ChIP-seq data analysis. After obligatory quality control and filtering of low quality, reads are mapped to the reference genome. Afterwards analysis calls for significantly enriched regions (peak calling) and these regions are mapped to genes (gene locations) in the genome (peak annotation). Figure 2.9 illustrates the process of ChIP-seq analysis, starting from raw reads up to extract significantly enriched genomic regions.

In chapter 6 ChIP-seq data is used to determine TF-binding events that play a crucial role during T helper cell differentiation. Another use for ChIP-seq data is transcription factor binding site (TFBS) or Motif discovery in the promotor region of investigated genes. Retrieved sequences are used to search for known (TFBS) or to identify unknown (Motif discovery) binding locations for TFs and small RNAs.

RNA sequencing

RNA sequencing (RNA-seq) refers to sequencing technology used to determine mRNA expression. In contrast to microarray technology, RNA-seq measures known and novel transcripts in coding and non-coding RNA. Quantification of RNA expression by sequencing is based on the number of detected sequences (reads) matching a specific coding region in the genome. The major advantage of RNA-seq towards microarray technology, is the ability of detecting so far unknown transcripts (transcripts that do not match any

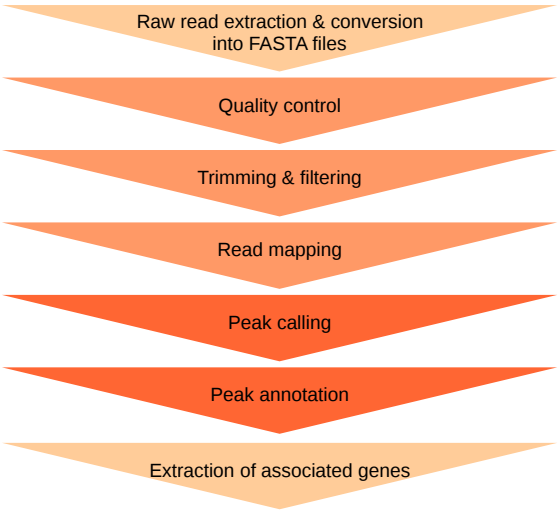


Figure 2.9.: **Chromatin Immunoprecipitation sequencing(ChIP-seq) peak calling analysis workflow.** Analysis workflow consist of six steps followed by an additional annotation step (extraction of associated genes). While step 1-4 are similar to RNA-seq analysis, peak calling and peak annotation are specific to ChIP-seq analysis.

known sequence). Thereby, sequencing allows to measure "all" expression in contrast to microarray which only measures mRNA sequences that are presented on the chip. "All" expressions also includes expected and unexpected changes of the genome, e.g. mutations, single nucleotide polyphormisms (SNPs), copy number variations (CNVs), insertion and deletions (InsDels) as well as the identification of new molecular entities like micro RNA (miRNA), smallRNAs [225].

In principle, RNA-seq allows analysis of all expressed transcripts, with three key goals: (i) annotating the structures of all transcribed genes including their 5'- and 3'-ends and all splice junctions, (ii) quantifying expression of each transcript and (iii) measuring the extent of alternative splicing [212]. A basic workflow of RNA-seq data analysis is illustrated in Figure 2.10.

In Chapter 6, we will focus on RNA-seq used to determine the expression profiles of Th2 cells during activation and differentiation from naïve T cell into Th2 cells. Thereby, we integrate different RNA-seq data sets for certain time points during the process and enable whole-transcriptome analysis to investigate TF-activity and estimate potential interactions via detection of co-expression of genes.

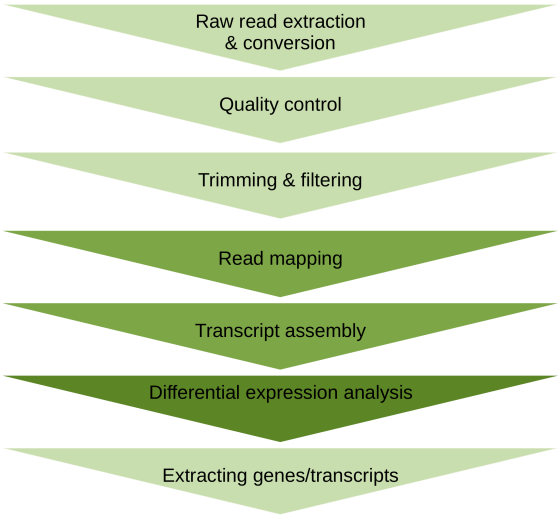


Figure 2.10.: **RNA sequencing(RNA-seq) expression analysis workflow.**The steps 1-4 are similar to the ChIP-seq analysis workflow (Figure 2.8), while step 5-7 are specific to RNA-seq analysis. The general outline of the workflow is based on the work of Trapnell et al. [359]

3. Meta-analysis of large gene expression experiments from public data repositories

In this chapter we discuss methods and applications of meta-analysis of gene expression experiments for aggregating single studies into a large meta experiment. Until today, high-throughput gene expression experiments are time and money consuming, especially when large amounts of data are necessary to analyze gene regulatory mechanisms. Meta-analysis can help to create large assembled experiments to overcome these shortcoming by using existing data. Furthermore, it has been shown that meta-analysis can increase statistical power of detecting differential expression [64] and significantly improves reproducibility when compared with independent studies [146].

The aim of this chapter is to demonstrate a novel meta-analysis approach to aggregate large data sets of Treg cell experiments for subsequent analysis of expression differences between Treg cell subtypes, namely iTreg and nTreg cells. We show how data was collected, curated and assembled to create a large gene expression data set. This includes quality checking and the assessment for batch- and lab-effects between data from different sources and technologies. The final data set is presented as a gene expression matrix containing over twelve thousand gene expression values retrieved from over 150 microarrays and more than 30 studies.

In the related work section we describe existing methods of meta-analysis for gene expression data. Afterwards we introduce our strategy of collecting and assembling gene expression experiments from public repositories (Section 3.4.1-3.4.3). Next, we describe our approach to process assembled data while focusing on reducing the experiment specific biases. Thereby we enable subsequent data analysis using established methods (Section 3.4.5-3.4.6). Finally, our findings are presented in the results and summary sections. We applied our strategy to Treg subtype specific experiments, as our goal is the assembly of a large meta expression matrix to enable the identification of distinct surface marker genes for iTreg and nTreg cells using machine learning, which will be covered more detailed in the subsequent Chapter 4.

The methods and results of both chapters were originally developed for investigating surface marker proteins for regulatory T cell subtype distinction, in cooperation with the German Rheumatism Research Center in Berlin

3.1. Motivation

The major aim of meta-analysis is to enable analysis of assembled experimental data from different source. Such assembling helps to overcome restriction regarding lower data amounts of specific analysis methods. Furthermore it enables the re-use of experimental data by assembling existing experiments instead of (re-)performing them, that is expensive, time consuming or simply spoken - inefficient, as this experiment data implicitly already exists. Thereby, meta-analysis must solve several issues, like handling variations (noise) in and between studies, without removing "real" signals, like differential expression or deciding whether studies are comparable or not.

Since high-throughput methods became more and more popular and the amount of available experiments increased and still grows, meta-analysis enables the (re-)use of existing data. Another need for meta-analysis emerged from the fact that up to a certain point the impact of time or condition changes cannot be avoided during lab-work due to simple constrains like sequential order or incremental setups of experiments or other technical limitations. This accounts for other technologies beside microarrays, too, e.g. RNA- or DNA- sequencing and even for technologies like histopathological images [196].

Two basic classes of meta-analysis methods were established, *merging raw data* and *combining existing results*. While merging raw data allows user full control over methods of data analysis, combining results releases user from normalizing, correcting and adjusting raw data in advance of further context specific analysis. For the field of merging raw data, two major fields of applications exist. The first is using existing data to create large data set to further analyze these using methods that have minimal requirements to the set size. For example some GRN methods that need at least as many experiments as regulators are requested. And second, the field of clinical applications, where patient data is permanently added to a cohort, which is analyze as one single experiments. Setups as the second are used for example to extract structural genomic anomalies, disease specific gene expression pattern, validate drug and changes of their targets [79, 292]. Another reason for the popularity of such approaches is the increasing availability of large curated, domain-specific data sets and hosting repositories

While the necessity of the first will slowly decrease in the next years, because high-throughput technology allow for cheaper and larger data generation (Figure 2.7), the second will keep valid since extraction of patient material and the influence of surrounding settings (like even room temperature or chronology of working steps) cannot be standardized over large periods of time.

3.1.1. Key issues for meta-analysis

As many approaches for meta-analysis with varying quality have been developed in the last decade (see Section 3.3), Ramasamy et al. stated seven key issues for meta-analysis [285]:

1. Identify suitable microarray studies,
2. Extract the data from studies,
3. Prepare the individual data sets,

4. Annotate the individual data sets,
5. Resolve the many-to-many relationship between probes and genes,
6. Combine the study-specific estimates,
7. Analyze, present, and interpret results.

The presented selection and manual curation in this work directly follows the first issue Ramasamy mentioned. Individual data set retrieval, quality checking, normalization and individual identifier (Id) mapping reflect the issues two to five. Issue six lead us directly to using a strategy were raw data is used as we plan to further analyze the data based on the expression values. Finally, issue seven is the object of Chapter 4 of this work.

3.2. Data source for building meta expression experiments

The basic requirement for the assembly of meta expression experiments is data. Back in the days, researchers often only published results without attaching the raw data that was generated and analyzed to achieve these. But, in the last 10-15 years more and more scientific journals and open access initiatives started to demand to publish the raw data that corresponded to scientific publications. A major intention was to increase traceability and reproducibility of scientific experiments and results. Nevertheless, this development also enabled the re-use of published data for meta-analysis, as the amount of available raw data of experiments increased permanently. Especially because detailed information about the experimental context and technical descriptions are often provided. The tremendous amount of published raw data strengthen the demand for central instances that enable access to the data. Consequently scientific online data repositories emerged.

3.2.1. Online data repository

Data retrieved from high-throughput gene expression experiments, mainly using microarrays, tremendously increased over the last 15 years and thus the potential for meta-analysis. In these years data repositories emerged to tackle and overcome the problem of data scarcity and permit the demand for researcher to publish their raw data. Multiple online accessible data repositories ease the acquisition of publicly available experimental data. NCBI's Gene Expression Omnibus data repository (GEO) and EMBL-EBI's ArrayExpress are the most popular and largest repositories for gene expression data. Up to date both platforms contain about 60-70 thousand experiments from array or sequencing experiments. Already in 2011 Barret at al. assessed the size and growth of data for the GEO repository over the last ten years exemplarily [22]. Two years later, the same authors described the tremendous amount of uploaded sample and the continuous exponential growth of the database [23]. Figure 3.1 illustrates the growth of the database between 2000 and 2016. Another advantage of GEO and ArrayExpress is that third-party libraries for scripting languages allow remote access and automatic retrieval of data and meta-information via Application programmer interfaces (APIs).

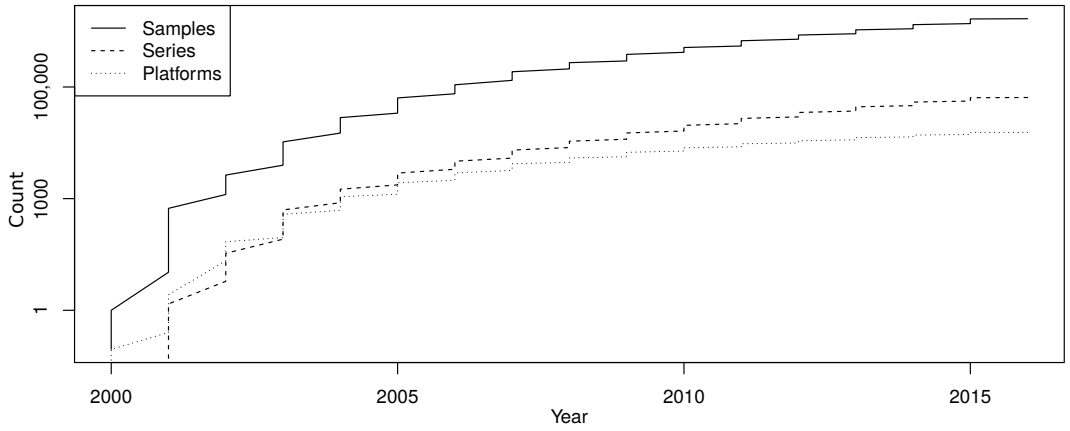


Figure 3.1.: **GEO Gene Omnibus platform growth over the last 15 years.** Plot illustrates the orders of magnitudes of platform, studies and samples submitted to the GEO database each year since inception. Until January 2016 about 1,669 million samples from 64,816 series (studies) perform on 15,358 were submitted to the GEO database. (Numbers were take from the website (`https://www.ncbi.nlm.nih.gov/geo/summary/?type=history`, online on 2016/01/10)

Over the years other, more organism, or tissue specific repositories emerge, but also disappeared, as for example The Arabidopsis Information Resource (TAIR) [154] or the Stanford Microarray Database (SMD) [116]. Finally both were integrated into GEO and ArrayExpress.

The advantage of many of these such "specialized" repositories or databases (e.g. M^2DB [62]) is, that many of these perform manual data curation, which means, that additional steps of quality control and meta-information annotation is applied before integrating data sets into the repository. Thereby, next to automated quality control, experts and scientists evaluate the data sets before integrating them into the repositories. Furthermore, data is enriched by additional information [8]. Although, many of these are commercially driven, they often grant access to researcher and academia, like Oncomine - an compendium of cancer transcriptome profiles that features additional analysis tools [291] or BloodSpot [16] a repository for cancer studies, that hosts data for normal and aberrant blood formation studies.

In the field of biological sciences various repositories covering different scope and aspects emerged. A number of repositories evolved that focus on hosting sequencing data. The limitation to sequencing data is caused by multiple reasons, one is the establishment NGS technology and the second the huge requirements on computational power and storage to host and analyze such data [282, 337]. Some of the most popular repositories are:

- DNA DataBank of Japan (DDBJ) [189]
- EMBL Nucleotide Sequence Database (ENA) [177, 339]

- GenBank [26]
- dbSNP [323]
- dbVar [262]
- EBI Metagenomics [159]
- NCBI Trace Archive [70]
- NCBI Sequence Read Archive (SRA) [70]
- IGSR: The International Genome Sample Resource (The 1000 genomes project) [67]
- The ENCODE (ENCyclopedia of DNA elements) project [69]
- RNASeqMetaDB [123]

The Immunological Genome Project [320] is an repository that aims to provide genomic data regarding the immune system of mice. As all integrated data is generated by the project members itself, analysis and tools for further investigation are provided. But, to this point the repository is rather small (816 arrays of 249 cell types, from 27 tissues in 2012) [320]. Other immuno-related repositories were not further maintained or integrated in others, like Macrophage Expression Atlas (GPX) [121] or Resource of Asian Primary Immunodeficiency Diseases (RAPID) [181]. Unfortunately, today only very few repository focusing on immune cells exist, the most popular project is the *Immunological Genome Project and systems immunology (ImmGen)*. ImmGen focuses on genes and their networks in the immune system, while it aims to construct regulatory computational models to identify known and novel regulatory interactions [320]. While ImmGen is an online platform to analyze and annotate data it also uses Gene Expression Omnibus (GEO) as data hosting platform. The access to preprocessed or normalized and categorized data is restricted. Therefore, we make use of search and filter functions of the GEO that contains many appropriate experiments as described in Section 3.4. Nevertheless, additional manual data curation is absolutely necessary to ensure that selected experiments meet specific requirements to assemble a data set that enables valid comparison between Treg cell subtypes and further analysis.

3.2.2. Data curation

When merging or assembling data from different sources one aspect of quality control is to assess the comparability. As we suggest to use large data repositories, that cover various kind of data unspecific to a certain field of research it is necessary to check wether selected data does in fact contain experiments, that meet the study aim. Therefore, manual curation is an additional preprocessing step that ensures this aspect of data quality. In 2001 Brazma et al. presented Minimum Information About a Microarray Experiment (MIAME) a standard that describes which minimal information must be provided to a microarray experiment [39]. This minimum of information should ensure the interpretation and the verifiability of experiments. Brazma et al. defined six section of MIAME that describe a microarray experiment:

1. Experimental design: the set of hybridization experiments as a whole
2. Array design: each array used and each element (spot, feature) on the array

3. Samples: samples used, extract preparation and labeling
4. Hybridizations: procedures and parameters
5. Measurements: images, quantification and specifications
6. Normalization controls: types, values and specifications

These features support the integration of new experiments into repositories and provide important information for reproducibility of experiments or for latter re-use, as for example for meta-analysis.

3.3. Related work

Up to date, various methods of using existing data to create new artificial data sets from microarray expression data exist. Already in 2002, Rhodes et al. aimed for stable cancer markers over different individuals, with varying experimental setting, conditions and patient states. They proposed meta-analysis of different patient studies to overcome the problem of data sparsity. Furthermore, Rhodes et al. assumed that meta-analysis produces more stable markers with increasing amount of data from different studies [290].

As early as 1999, Lander [200] reviewed the application and the resulting opportunities for microarrays and the advantages of combining multiple studies to enable global expression and genome analysis. As one of the first, Choi et al. analyzed batch effects as an issue of systematical differences between studies [64]. This work was a follow-up study to the meta-analysis study of Rhodes et al., who aimed to identify prostate cancer related genes from multiple microarray studies [290]. In the following years various methods to reduce the batch-effect were presented [10, 25]. But, as Johnson et al. wrote in [171] many of these methods required large amounts of data to work properly and to not diminish biological signal. Various methods have been developed in the last 15 years using different approaches, e.g. single values decomposition [10] or empirical Bayes methods [171, 207, 251].

3.3.1. Comparison of methods for meta-analysis of gene expression data

In 2007 Cahan et al. [47] distinguished between two basic classes of microarray meta-analysis. The first class depicts merging raw data from different studies and considering the merged data as one experiment in subsequent analysis. In the second class experiments that are selected for a meta-analysis are provided as individual lists of differentially expressed (differentially expressed) genes or list of summarized gene expression changes. It is important to note, that thereby individual studies were not necessarily analyzed using same or even compatible methods. The first approach is widely used when patient data from different studies is aggregated [226, 375, 387]. An example of the second class is the integration results from multiple studies to develop gene interaction networks from different experiments and even from different technologies (microarray, RNAseq, DNAseq) as described in [168]. Both classes exhibit certain advantages (+) and disadvantages (-) regarding different aspect as these are:

Merging raw data	Combining result lists
- High computational and technical requirements	+ Low computational and technical requirements, no raw data processing
+ Fine-grained selection of experiments (parts of studies)	+ Integration of multiple technologies, e.g. microarray, RNAseq
+ High comparability of results	+ Integration of heterogeneous source and formats
+ Reproducibility of complete analysis pipeline	+ No preprocessing effort
- High effort on integration of different techniques and platforms	+ Extensibility by additional data, without reprocessing data
+ Quality control of raw data	- No quality control of raw data
+ Batch- /lab-effect estimation	- Compatibility of processing methods required
+ Re-computation using different (new) methods	- Ambiguity of mapping – compatibility required, e.g. ProbeSets to transcript to gene
+ Freedom of choosing methods for subsequent expression analysis	- Comparability of results restricted to applied processing and analysis methods in original experiments
+ Flexible ID mapping, e.g. using different reference genomes	

An example for the concept of comparing results is used by GEO Gene Omnibus online portal, that does provide web features such as, search and on demand visualization of individual gene expression data.

Along with the individual advantages come distinct fields of application of both classes. On the one hand biologist tend to compare results from different publications, where on the other hand bioinformaticians or statisticians often prefer raw data to apply their favorite methods, existing frameworks or scientific workflows. Thereby, biologist do not need specific skills for data processing and can focus on results from different studies as these may also include varying technologies for gene expression detection, like RNA-seq and microarray experiments. Furthermore required computational power to process and analyze large amounts of raw data prior integration into a meta set must be available.

Another reason to perform meta-analysis by merging raw data is to overcome batch effects, where data/material for different samples is processed time and even location (lab) independently [295]. In 2012 Tseng et al. prepared an overview on existing methods of meta-analysis for gene expression experiments [361].

3.3.2. Types of meta-analysis

Choosing a particular type of meta-analysis method is mainly driven by the overall goal of a study. Beside differential expression analysis, also gene function prediction, target gene detection, and gene cluster detection demand for different analysis strategies. For instance, differential expression analysis compares gene expression of two samples from

different batches, where a conceivable aim of a gene cluster analysis is the identification of homogeneously expressed genes, which does not require a direct comparison between samples or states and therefore batch effects are negligible. In consequence, choosing appropriate strategies of meta-analysis depends on the available input data and the aim of the subsequent analysis.

Before presenting implementations in detail, we give an overview of meta-analysis methods. Five of the most relevant methods of meta-analysis are the following (based on [361]):

P-value combination allows for integration of results from different studies and belongs to the second class. For example Rhodes et al. suggest to combine P-values by summing up minus log-transformed P-values for a gene over all studies. The received score indicates the strength of differential expression [290]. Next to other transformation methods, some approaches calculate min or max P-values over all and use these as test statistics to estimate significance of differential expression [217]. Obviously, only studies sharing the identical or very similar aim should be combined.

Combining effect size is based on the idea that each gene expression profile for each study is the combination of the true effect size (real gene expression changes) and a study- or batch-specific error. The method belongs to the first class as it combines raw data. Different authors approach to model the effect using fixed effect models (FEM) or random effect models (REM) between studies. FEM assume the effect size to be identical for all integrated studies [37, 311] where REM try to cover heterogeneity of studies caused by sample or processing differences, which is much more realistic [64].

Combining ranks denote the use of ranking results (e.g. differentially expressed genes), which are combined using robust rank statistics [214, 409]. Outliers are a constant problem of such methods, when rank aggregation is performed using mean or similar statistics [361]. Approaches like RankProd try to overcome such problems by calculating the product of ranks of fold changes (fold changes) instead of P-values in each inter-pair group of samples [147]. As the fold changes are not recalculated, combining ranks is part of the first class methods.

Directly merging raw data is the class of methods that combine raw data as a single study before performing further analysis like normalization. As such methods do not involve inter-study effects (also called effect-size) handling, additional effort is necessary to account for batch-effects or errors [326]. Direct merge of raw data is often applied when studies use identical microarray technologies or even identical microarray types [205]. In recent years, several methods and R packages have been developed to enable direct merging of raw data with subsequent batch-effect handling as e.g. BMC [25], COMBAT [171], GENENORM, and XPN [317, 347].

Latent variable approach describes a probabilistic model to integrate studies. The basic idea is to use a latent variable to model quantities of gene expression that can be

combined over studies. For example, Choi et al. used latent variables to estimate an index – the probability of expression (POE), that is used to compare two groups of a meta data set [63]. This approach is part of the class of methods that merge raw data.

3.3.3. Existing meta-analysis approaches and implementations

A major issue on meta-analysis is the problem of removing signals, that occurs during each adaption, transformation or shift of expression values up to a certain amount, regardless of the used method. The better the method estimates differences based on batch- or lab specificity the lower the loss of "real" differences. Several authors discussed this problem coming along with meta-analysis starting from study selection up to mapping strategies between different platforms [48, 266, 285].

As early as 2002, Rhodes et al. [290] applied a meta-analysis strategy on different prostate cancer microarray experiments with the idea of identifying recurring gene expression pattern across different studies and technologies. They found a set of genes that was consistently and significantly dysregulated across all experiments. Rhodes et al. did not account for inter-study effects as they disregarded the actual strength of dysregulation but considered shared differentially expressed genes between different data sets.

In 2003, Choi et al. [64] extended the idea of combining data sets with a systematic approach to handle inter-study differences. The approach summarizes the individual bias of an experiment, the so called effect size. The effect size implements a statistical framework for combining different results by comparing intra- and inter-study differences of expression levels for single genes under consideration of the inter-study effect, that is measured by comparing the average measure of differential expression across all data sets for each gene. The effect size was estimated via a fixed effect model (for limited number of experiments performed under accountable conditions), a random effect model (for published experiments from different origins) and by an Bayesian approach, that offers more flexibility and robustness for varying inter-study differences.

Sirbu et al. [346] proposed a method for meta-analysis in preparation of gene regulatory network reconstruction from public data. They focus on integration of single- and multi-channel arrays and compare their method against Johnson's ComBat method [171] and the XPN method published by Shabalin et al. [317]. Before normalization a preprocessing step is performed, where values were estimated with a mixed method of Loess normalization [330] (for Affymetrix single-channel arrays) and PM-only normalization (for dual channel arrays) [332]. Sirbu et al. normalized cross-platform via expression values transformation to values between 0 and 1 after preprocessing each data set by subtracting sample mean and division by sampled standard deviation of the sample. The transformation was done by subtracting the overall smallest expression values from all values and subsequent division by the greatest values in the data set [346].

Johnson et al. [171] implemented an empirical Bayes approach (ComBat), which consists of (i) standardizing expression values per gene to similar means and variances, (ii) batch effect parameter estimation using parametric empirical priors and (iii) the adjust-

ment of the data to the batch effect. The authors note that before estimating the batch effect all genes that were "absent" in at least 80% of the measurements were removed to eliminate noise.

XPN (cross platform normalization) method [317] enables normalization of studies from different array types. After pre-processing the data and reducing the experiments to a common gene/probe set each experiment is represented as a matrix and then combined into one combined matrix. Via multiple k-means clustering XPN identifies homogenous groups of genes and samples in combined data. For each block the specific values are shifted to its mean plus noise. To achieve robustness and minimize effect of initial cluster centroid selection the results of multiple clustering iterations are averaged and assigned.

Frozen robust multiarray analysis (fRMA) by McCall et al. [251] approaches the challenge of removing batch effects, where a batch is defined as a number of microarrays that are processed at same time and lab, while an experiment consists of multiple batches that use the same technology (array type). fRMA is very similar to the RMA method for single experiments but does account for differences for the batches by calculating a reference distribution from a database of publicly available expression data. After background correction all single arrays are (quantile) normalized regarding the reference distribution. Finally expression values are summarized by correcting the expression value by the calculated batch effect.

Conlon et al. compare two Bayesian approaches [68] that belong to the class of latent-variable approaches. One that combines standardized gene expression measures across studies, while the second combines probabilities of differential expression. Later approaches like GeneMeta [239] and its extension MetaArray [114] focus on comparing expression profiles and extracting differentially expressed genes between different data sets that were processed on identical microarray platforms (GeneMeta). Therefore a user must i) perform a gene-id mapping before applying MetaArray if different platforms are used, and ii) assess the comparability between different experiments.

Campain et al. extended Yangs Differential Expression via Distance Synthesis (DEDS) [392] to enable meta-analysis (mDEDS) [48]. It makes use of multiple statistical measures to obtain a list of differentially expressed genes. In contrast to other methods like MetaArray, differentially expressed genes are detected regardless of used platform types. Similar to RankProd, mDEDS assumes that relevant genes can be detected between multiple experiments by rank combination. As the introduced method performs the actual aggregation after detecting differential expression it belongs to the class of rank combining methods.

MergeMaid by Cope et al. [71] provides function to merge different expression sets into a single one. MergeMaid enables visual comparison of specific samples or probes (e.g. genes) between different studies, but requires user-supplied mappings. MergeMaid alone does not provide mechanism to handle or overcome differences between array types, or batch effects. Therefore the subsequent application of a batch effect reduction method is necessary.

Another field is the problem of mapping identifier from different platforms. Anova Sibling Consolidation by Li et al. [216] is a method to test comparability between experiments from different microarray platforms. They utilize *Anova* test statistic to check

whether expression profiles of two probes mapping to the same gene are similar expressed and thus can be merged or not. This approach can be adapted to check same expression for two probes from different experiments or even different microarray platforms as long as a mapping between them is available.

Comparison of methods Evaluation of different methods (not identical to prior presented) revealed that there is no overall best working method, rather then different methods have perform better regarding certain criteria. To shed some light onto this, Chang et al. reviewed and compare 12 meta-analysis methods [56]. They formulated four criteria, namely (i) detection capability, (ii) biological association, (iii) stability and (iv) robustness.

Detection capability is an approach to assess how many differentially expressed genes are detected under the same assumptions regarding significance for all methods.

Biological association denotes the results of pathway enrichment analysis and the relation to the targeted disease, which is implied by the underlying data set.

Stability is detected by splitting data sets into subsets and compare results of meta-analysis methods regarding overlapping differentially expressed genes.

Robustness of methods is determined by adding outlying studies to the data sets and measuring their influence towards the prior detected set of differentially expressed genes and the relative difference.

An important conclusion of Chang et al. was, that the most appropriate method depends on study goal and the underlying data set, as robustness and stability might be better covered by methods that did not performed best over all criteria.

3.4. Assembling gene expression experiments for meta-analysis

As mentioned in the motivation our approach is guided by the key issues for meta-analysis by Ramasamy et al. [285]. Figure 3.2 outlines our strategy starting from data collection until extracting the final meta expression matrix. Here, we perform meta-analysis by assembling a large curated Treg cell gene expression compendium from public data. W e describe each step in detail, followed by the results presentation in Section 3.5.

3.4.1. Data collection

To collect data we used NCBI's GEO web portal (`http://www.ncbi.nlm.nih.gov/geo`) and the query *T regulatory cell AND mus musculus*. We obtained a list of 1500 arrays. GEO organizes microarray and other high-throughput experiments by four different categories: GEO platform identifier (GPL), GEO sample record (GSM), GEO series record (GSE), and GEO dataset record (GDS) [363]. The GPL specifies a concrete technology the experiment is based on, e.g. the specific microarray chip or sequencing

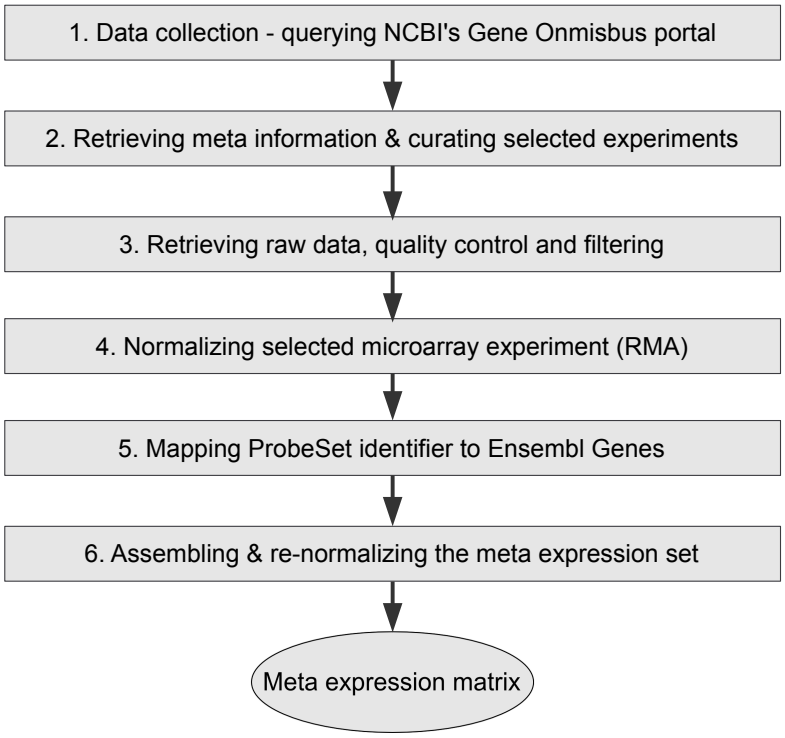


Figure 3.2.: **Schema illustrates the steps of meta-analysis starting from requesting online repositories up to assembling the meta expression matrix.** Sections 3.4.1 - 3.4.6 describe the pipeline in detail.

technology that was used to perform the experiment. GPLs are unique up to the manufacturers version of a technology like different versions of a specific microarray chip. The GSM identifies the specific microarrays record and its output files. Each GSM is unqiue over the complete set of all experiments in the GEO repository. A GSE can specify all experiments that were processed in one study or, more abstract, all experiments that share a common context, not necessarily processed together in a single study. Experiments (GSMs) can be associated to more than one study and each study can inherit experiments from others. Thus, meta studies can be modeled as it is presented in this work.

3.4.2. Retrieving meta information

The list of extracted experiment Ids was filtered for study type *Expression profiling by array* using GEO webportal filter tools. In the next step meta information and description for each experiment were retrieved using R package GEOmetadb [408]. All information about the GEO experiments are stored in a SQLite database, which can be accessed locally using structured query language (SQL). Thus it is possible to access the information via introduced identifers like GPL, GSM or GSE.

We created a list containing title, organism, GSE, GSM, GPL and the description of all filtered experiments to enable immunologist to manually identify and remove experiments that do not satisfy a defined set of minimal requirements. Next to organism information immunologist focused on the cell type mentioned in the description, the stimulus, and experimental details like viral treatments, knock-out (KO) or knock-in (KI) information. Table 3.1 contains a detailed list of the major criteria used to include or reject experiments to our collection.

Table 3.1.: **Listing of criteria for inclusion or rejection of experiments while manual curation by immunologist.**

Include criteria:	**Exclude criteria:**
• In-vitro and in-vivo induced iTreg cells with activation period longer than 3 days • Fresh ex-vivo and activated nTreg cells from wild-type mice, Balbc mice (lab mice), C57BL/6J (lab mice), NOD mice, TCR-HA transgenic mice, Foxp3 reporter mice	• In-vitro and in-vivo iTreg cells induced for short-time only (3 days or shorter) • Genetically modified mice with deficient Treg cell master transcription factor Foxp3 • Cell mix of iTreg and nTreg cells • Scurfy mice • Genetically modified mice with deficiency in Il-2 pathway

Furthermore time information was added if mentioned in the specific experiment or study to ease the subsequent manual categorization into nTreg or iTreg cell experiments. This is necessary as the stated aim of the resulting meta expression set is to identify distinct surface marker gene for both Treg cell subtypes. The process of manual curation resulted in a set of 300 microarray experiments assigned to 74 different studies (GSE), where 48 were initially classified as nTreg cell and 26 as iTreg cell related studies. Due to constraints of our normalization and aggregation pipeline the data set was further limited to experiments performed using Affymetrix gene chip technology (see Figure 3.3). Table 3.2 summarizes the initial amounts of experiments returned by GEO Omnibus web portal upon request and the finally selected set of 154 microarrays.

3.4.3. Retrieving raw data

In the next step the raw data, here Affymetrix CEL files (containing the data extracted from an Affymetrix GeneChip) were retrieved using R-package GEOquery [80]. GEOquery enables the automated download of data by GSM/GSE accession numbers from the GEO repository.

Studies or single experiments were deselected if either no experimental raw data were available or if the data did not pass a quality check using arrayQualityMetrics [180], i.e. were detected as outlier in three or more tests. This criteria led to the exclusion of five experiments. It is important to mention that quality control tests were applied to each study (GSE) in isolation to prevent outlier-detection coming from differences between studies (see Section 3.4.6).

All these selection and filtering steps led to a final set of 154 arrays from 36 studies,

Table 3.2.: **Table shows the size of the collected data set for each step of the pipeline.** The column "matched GEO-query" contains the numbers of elements as retrieved from GEO, while "selected" contains those, included in the meta expression data set after filtering and quality control.

	matched GEO-query	selected
Arrays	604	154
Data sets	238	36
Treatments	-	15
Platforms	51	6
Species	1	1
Cell types	>4	2

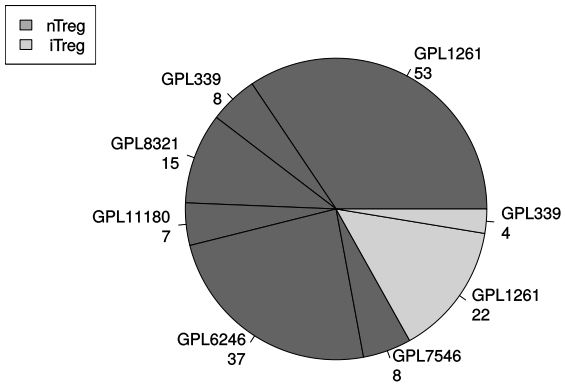


Figure 3.3.: **Fractions of platform types in the final meta expression set.** 154 arrays (29 iTreg and 125 nTreg) from 36 GEO series were selected. This set contained six different Affymetrix GeneChip®technologies. Numbers below the platform type indicate the corresponding amount of integrated microarrays.

where 29 arrays originated from iTreg cells and 125 arrays from nTreg cells (see Figure 3.3). Appendix A, Table A.1 contains all selected microarray experiments that were integrated in the meta expression set.

3.4.4. Meta expression set

After collecting the data as described in the previous steps, data must be preprocessed to allow for a joined analysis. Due to the heterogeneity of the experimental settings under which the data was obtained, a proper normalization of expression values is of highest importance. Such a normalization must take intra-study and inter-study variety into account. While intra-study variety relates to the general problem of normalizing measured expression intensities between microarrays [332], across-study variety refers to multiple effects and systematic differences between experiments that were not part of

the same study [47, 311]. To overcome the problems of intra- and inter-study effects a two-step algorithm was implemented resulting in comparable expression values over the complete set of selected studies.

In the first step, all measurements within the same study were normalized using RMA (robust multi-array average) published in 2003 by Irizarry et al. [163] and provided by different R-packages like the Affy package [110, 283]. Required study-specific information about replicates, groups of arrays, temporal or circadian information, so called phenodata were taken from previously retrieved meta data. The phenodata of a study basically includes the description of each array and its association to a specific subgroup that share a certain property, e.g. cell type, stimulation, stimulation time, treatment and tissue. In the second step, we reassigned the selected microarrays in either the group of iTreg cell or in the group of nTreg cell experiments. Disregarding small differences in sample preparation or similar we consider such differences as biological variance. The resulting set of experiments from different studies, assigned to one of both groups was necessary to enable subsequent assembly of a large expression set. The next step was the definition of a common set of gene Ids over all studies to enable meta set assembly.

3.4.5. Identifier mapping

Next, we built a single meta expression matrix across all studies including the intersecting of ProbeSet Ids of all experiments. Thereby we mapped original ProbeSet Ids to Ensembl Mouse Gene Identifiers (ENSMUSGs) using biomaRt [89]. In detail we retrieved all ProbeSet Ids to ENSMUSG mappings for each of the six used microarray types (Tab. 3.3). As multiple ProbeSets on a standard Affymetrix microarray are related to transcripts (to cover different exons and transcripts of genes) and thus can be part of the same gene. To solve this problem a surjective mapping from multiple ProbeSet Ids to one ENSMUSG is required.

Table 3.3.: **Overview of integrated gene expression arrays.** All microarrays are in situ oligonucleotide arrays manufactured by Affymetrix.

GEO Accession	Title
GPL339	Affymetrix Mouse Expression 430A Array [MOE430A]
GPL8321	Affymetrix Mouse Genome 430A 2.0 Array [Mouse430A_2]
GPL1261	Affymetrix Mouse Genome 430 2.0 Array [Mouse430_2]
GPL7546	Affymetrix GeneChip Mouse Genome 430 2.0 Array [CDF: Mm_ENTREZG_10]
GPL6246	Affymetrix Mouse Gene 1.0 ST Array [MoGene-1_0-st] [transcript (gene) version]
GPL11180	Affymetrix HT MG-430 PM Array Plate [HT_MG-430_PM]

As Affymetrix provides microarray platform specific annotations by chip definition files (CDFs) to handle transcript mapping, these annotations do not prevent the described mapping problem. A CDF specifies which genomic region is represented by one or more probes and which probes of a microarray form a ProbeSet. Here, we applied chip type specific CDF environments provided by Bioconductor platform for R [354]. Afterwards, biomaRt service [88, 89] was used to retrieved corresponding ENSMUSGs.

Over the years different strategies were published to handle multi-referenced genes, like [216, 237, 333]. Next to trivial approaches like "first come first serve", where the first matched Id and the corresponding gene expression profile are matched, more comprehensive methods like Anova Sibling Consolidation (ASC) [216] came up. Thereby, statistical metrics are used to determine if the expression profile of multiple ProbeSets with the same target gene can be combined or not. ASC does not suggest how to perform mapping if expression profile do not match [216]. Other meta-analysis methods handover this problem to the user or, like frozenRMA [251], restrict the type of platforms to ensure common Id sets as workaround. Since the selection of experiments in this work contains experiments from different platforms, we did not apply such method (for more details see discussion in Section 3.4.5).

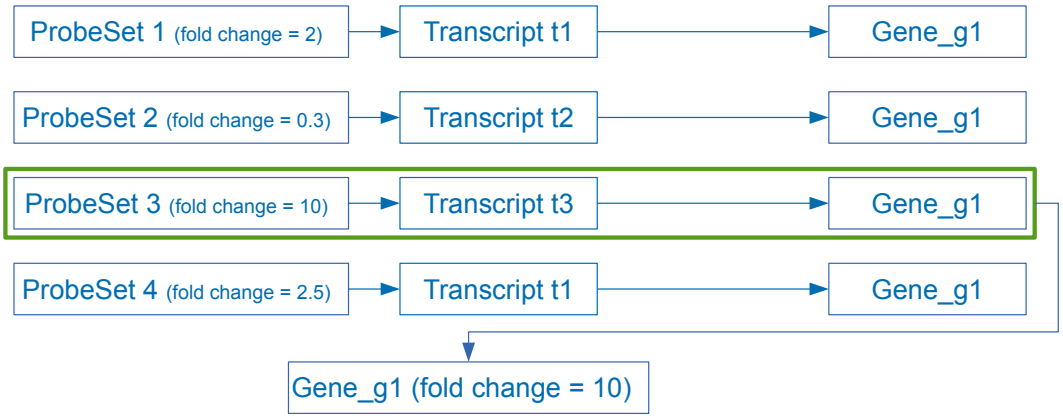


Figure 3.4.: **Different ProbeSet Ids link to the same Ensembl Mouse Gene Identifier** This problem is solved by assigning the expression values of the ProbeSet with the highest fold change between measured experimental states,(e.g. stimulation, cell type, treatment or time point) to the gene.

To solve the consolidation problem of assigning multiple expression values to a single ENSMUSG we implemented a simple straight forward way of mapping, where the ProbesSet with the highest intra-study expression change is selected and its expression values are mapped to the corresponding ENSMUSG (illustrated by Figure 3.4).

The mapping was performed for all 36 retrieved and normalized studies. After mapping all studies to a common set of ENSMUSGs expression data from all studies can be represented in one matrix (ENSMUSGs as rows and microarrays as columns).

3.4.6. Renormalization of assembled meta experiment

Now, we focus on the reduction of experiment specific expression pattern, so called batch and lab effects (also called effect-size or inter-study effect), that occur when experiments are performed under different conditions, e.g. different protocols, different laboratories or using samples from different individuals that are considered as similar. To reduce such effects several methods have been developed, like differential expression analysis or

Gene Wide Association Studies (GWASs). In the following, we describe our strategy to overcome such problems, using the prior introduced data set.

Lab-/batch effect

The batch effect on the assembled data set is visualized in Figure 3.5 and 3.6. To illustrate the effect we clustered the assembled data set for groups regarding the global expression profile over all genes (Figure 3.5). The average Euclidean distance between arrays of the same experiment over all experiments is 125.595, where the average pairwise distance between all arrays is 211.041. Figure 3.5 shows the hierarchical clustering of all arrays in the data set by expression values.

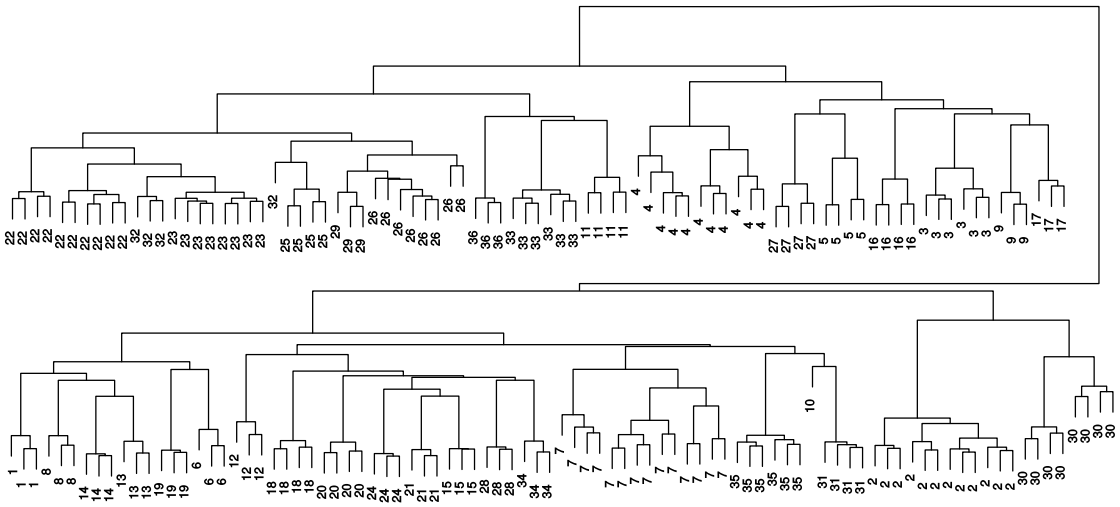


Figure 3.5.: **Hierarchical clustering of assembled meta expression matrix.** Most of the microarrays from the same experiment (GSE) cluster together, which illustrate the lab-/batch effect, as their expression profile is more similar among studies than towards other experiments.

Re-normalization strategy

The assembled meta expression matrix contains expression values of 12825 ENSMUSGs (rows) coming to 154 microarrays (columns) from 36 different experiments. The matrix reveals high variance for each gene. We performed an additional re-normalization step using quantile normalization to adjust expression values for inter-study effects. The quantile normalization transforms expression values for all single microarrays in the assembled data set to a value range fixed for all integrated experiments (see Figure 3.6). The idea is that the rank-based re-assignment of expression values as quantile transformation does not change the ranking of genes by expression values within an experiment, while it enables the direct comparison between studies. Such rank-based transformation allows for identification of expression change for a single gene by comparing its ranks for

different groups of arrays, independently of the specific value distribution of the single arrays or studies.

Table 3.4.: **Table illustrates expression value adjustment via re-normalization to improve comparability between data from different studies.** Table shows the means of all variances obtained for all genes in the complete data set and in the specific subsets (columns). The first row shows the varying distribution of expression values between assembled, but not re-normalized studies (GSEs). The second row shows the variances based on re-normalized expression values.

	Mean of variances within studies (GSEs) all cell types	Mean of expression variances within assembled meta experiment		
		all cell types	nTreg	iTreg
not adjusted	0.1221	1.8510	1.8710	1.3990
adjusted	0.1142	1.3030	1.2800	1.0840

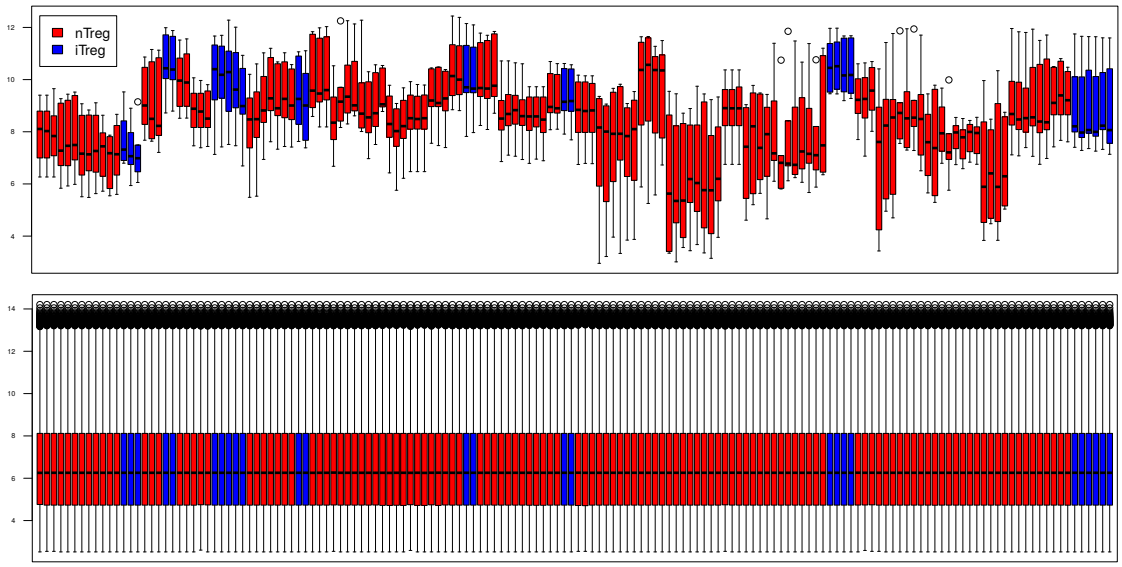


Figure 3.6.: **Upper boxplots show expression distribution of non-adjusted expression values after meta experiment assembly.** Varying expression values are clearly visible for both cell types (nTreg - red, iTreg - blue) and for all included studies. **Lower boxplots** show expression distribution after applying quantile re-normalization. Mean expression as well as minimal and maximal expression are adjusted to common levels, to enable subsequent expression analysis.

3.5. Results

In the following section we describe the results of data set retrieval, assembly, and re-normalization. Furthermore, we use differential expression analysis to evaluate the presented approach by detecting differentially expressed genes.

3.5.1. Data set assembly and re-normalization

After creating a large meta expression data set containing 154 nTreg and iTreg cell microarrays from 36 different studies, the data was preprocessed (quality control, Id mapping), normalized, and re-normalized for further analysis. The resulting meta-expression matrix consists of 12825 genes and 154 measurements.

Renormalization was performed to adopt for batch-effects between microarrays from different studies and reduced inter-study differences between the expression profiles. Also, re-normalization led to identical expression values for genes with the same ranks in different experiments and thus to identical value distributions, as shown in Figure 3.6. Table 3.4 illustrates this effect using mean variance and mean expression measures. While the variance of genes within an experiment does only decrease slightly, the inter-study differences of expression means decreased by about 20-30%.

Figure 3.7 illustrates the effect for four pairs of experiments. We randomly selected 200 genes out of 12825 to show the effect of re-normalization. The reduction of the gene set is necessary for better visualization. The microarrays used for one-to-one comparison in the Figure 3.7 were not part of the same study. The quantile-quantile-plots (QQplots) illustrate distribution shifts of these 200 randomly selected genes, where individual expression differences between genes in both experiments remain.

3.5.2. Differential expression analysis

One advantage of creating meta-expression experiments based on raw values is the opportunity to apply typical gene expression analysis methods, like differential expression analysis using t-test, GWAS or GSEA. Here, we present the results from differential expression analysis using t-test between iTreg and nTreg cell. Differential expression was calculated using R-package limma [297]. Genes were detected as differentially expressed when their absolute fold change was above 1 and the corresponding Benjamini-Hochberg corrected P-value below 0.05. We detected 877 genes to be lower expressed in iTreg cells and 260 genes to be higher expressed than in nTreg cells. Figure 3.8 shows expression differences for all differentially expressed genes between iTreg and nTreg. Out of the set of 12825 genes, 1137 genes were detected as differentially expressed before re-normalization and 891 of these were also detected as differentially expressed after re-normalization. Enrichment analysis of the non-overlapping genes of both data set states (before and after re-normalization) using DAVID [155] did not reveal any significant genes groups regarding Gene Ontology (GO) terms or pathways associated to Treg differentiation or Treg subtype fate. The set of 246 genes exclusively detected as differentially expressed prior re-normalization contains T cell specific genes. Additional, we analyzed the subset of

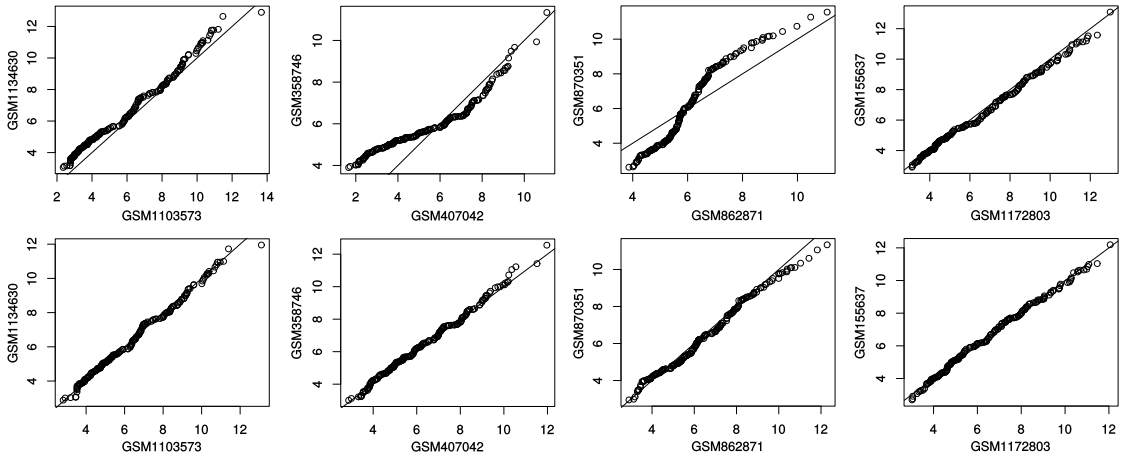


Figure 3.7.: **QQplots illustrate difference before and after re-normalization. First row of figures** shows qqplots of non-adjusted data, where **second row of figures** shows re-normalized data. By comparing upper and lower figures it becomes clear that expression values became more homogeneous, which illustrates the reduction of the inter-study effect (the shift onto the diagonal), at the same time the profile of expression differences between single genes are widely kept (compressions and stretches in line profiles). 200 randomly selected genes were used to plot four comparisons of expression profiles from two independent microarrays (origin from different studies) each time.

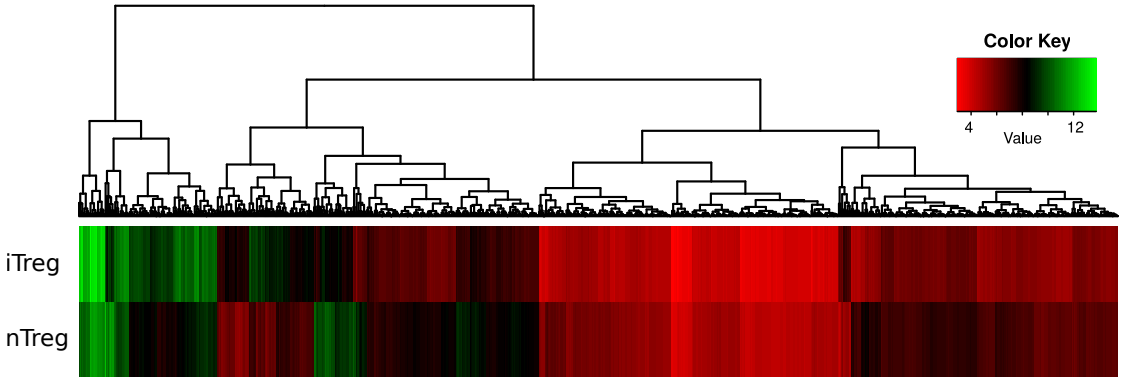


Figure 3.8.: **Genes differentially expressed between iTreg and nTreg cells.** Heatmap shows the mean gene expression extracted from the assembled and re-normalized meta expression set. Color key indicates expression values. Upper band of the heatmap represents expression in iTreg cells, where lower band represents expression in nTreg cells.

TFs from differentially expressed genes (Table 3.5) based on a published list of TFs by Wei et al.[378]. Except for *Ikaros*, non of these is known to be specific for nTreg or iTreg

cell, while some are Th cell related. For example *Jund, Il2* and *Tcf7* are known to be important during T cell fate decision [252, 301].

Table 3.5.: **Listing of differentially expressed TFs before re-normalization but not differentially expressed afterwards.** Except for Ikarus (italic written), all listed TFs are unknown to have a subtype specific gene expression profile.

Ensembl gene identifier	MGI gene symbol	Gene name
ENSMUSG00000014603	Alx3	aristaless-like homeobox 3
ENSMUSG00000035277	Arx	aristaless related homeobox
ENSMUSG00000039910	Cited2	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2
ENSMUSG00000026436	Elk4	ELK4, member of ETS oncogene family
ENSMUSG00000022101	Fgf17	fibroblast growth factor 17
ENSMUSG00000015627	Gata5	GATA binding protein 5
ENSMUSG00000002578	*Ikzf4*	*IKAROS family zinc finger 4; similar to helios*
ENSMUSG00000027720	Il2	IL2 (interleukin 2; T-cell growth factor)
ENSMUSG00000027947	Il6ra	interleukin 6 receptor, alpha
ENSMUSG00000071076	Jund	Jun proto-oncogene related gene d
ENSMUSG00000027985	Lef1	lymphoid enhancer binding factor 1
ENSMUSG00000038909	Myst2	MYST histone acetyltransferase 2
ENSMUSG00000029832	Nfe2l3	nuclear factor, erythroid derived 2, like 3
ENSMUSG00000028019	Pdgfc	platelet-derived growth factor, C polypeptide
ENSMUSG00000045302	Preb	prolactin regulatory element binding
ENSMUSG00000000782	Tcf7	transcription factor 7, T-cell specific
ENSMUSG00000022015	Tnfsf11	tumor necrosis factor (ligand) superfamily, member 11
ENSMUSG00000020923	Ubtf	upstream binding transcription factor, RNA polymerase I
ENSMUSG00000021514	Zfp369	zinc finger protein 369

Further gene set enrichment of the 891 differentially expressed genes using GO terms high ranks several T cell and differentiation related terms. In Table 3.6 we focus on terms that are highly specific for Lymphocytes and T cells, where terms that are related to general differentiation processes were left out.

The results from differential expression analysis and GO term enrichment analysis indicate, that re-normalization did not tremendously change the set of signals. The re-normalized meta expression includes 77% of the differentially expressed genes detected before re-normalization. This set also features Treg cell specific TFs (shown in Table 3.5). Furthermore, changes in the set of differentially expressed genes did not influence results from GO enrichment analysis, indicating that the overall expression characteristics were kept, while inter-study effects could be reduced (Figure 3.6 and 3.7).

3.6. Summary

In this chapter we described our approach of collecting and assembling gene expression microarray data to analyze Treg subtype specifics. Therefore, we constructed a large meta expression set based on data from public online repositories. We focused on data

Table 3.6.: **GO enrichment analysis of differentially expressed genes.** Using the enrichment online tool DAVID [155] (online,12/2015). 16 T cell related GO terms where identified as significantly overrepresented in the gene set. Terms are sorted from lowest to highest P-value.

GO ID	GO term	P-value
GO:0046649	lymphocyte activation	$8.66 * 10^{-6}$
GO:0046651	lymphocyte proliferation	$1.36 * 10^{-4}$
GO:0042110	T cell activation	$1.64 * 10^{-4}$
GO:0045596	negative regulation of cell differentiation	$2.03 * 10^{-4}$
GO:0002520	immune system development	$3.56 * 10^{-4}$
GO:0042098	T cell proliferation	0.0010154699
GO:0050670	regulation of lymphocyte proliferation	0.0026163637
GO:0030098	lymphocyte differentiation	0.0030818036
GO:0045597	positive regulation of cell differentiation	0.0038178296
GO:0050863	regulation of T cell activation	0.0047773595
GO:0042129	regulation of T cell proliferation	0.0077358778
GO:0048538	thymus development	0.0083282627
GO:0050671	positive regulation of lymphocyte proliferation	0.0088422211
GO:0051249	regulation of lymphocyte activation	0.0117505093
GO:0050868	negative regulation of T cell activation	0.0367450441
GO:0002711	positive regulation of T cell mediated immunity	0.0384057351

processed using Affymetrix GeneChip technology and studies that deal with regulatory Treg cells. To analyze Treg subtype specifics we searched and retrieved experiments that either performed expression analysis on induced regulatory T cells or natural derived regulatory T cells from mice. We applied quality control and did manual curation regarding cell extraction methods, treatments and stimulation.

After revising the list of experiments, we performed a two-step normalization to account for differences coming from heterogeneous condition experiments. The two-step re-normalization based on quantile normalization realigned the values distribution between the different studies and thus improved comparability.

Together, we extracted a gene expression set of 12825 genes and 154 experiments that allow gene expression analysis of iTreg and nTreg cells. Validity of our approach was assessed using GO term enrichment analysis and extraction and association analysis of 1137 differentially expressed genes. We examined the list of genes regarding known transcription factor involved in activation and differentiation of T cells.

At this point, the extracted and validated meta-expression set can applied to further analysis regarding the extraction of Treg subtype specific surface marker genes (Chapter 4) or the application of GRN reconstruction methods (Chapter 5).

Table 3.7.: **Enriched GO terms from categories Biological Process (BP) and Molecular Function (MF) for the comparison of differentially expressed genes detected before and after re-normalization.** Enrichment was performed using DAVID Tool [155] (online 12/2015). The set of significantly enriched terms (P-value < 0.05) does not include T cell or lymphocyte specific terms. Terms below the line are not significantly enriched, but enriched

GO ID	GO Term	P-value
GO:0022403 (BP)	cell cycle phase	$8.07*10^{-7}$
GO:0000279 (BP)	M phase	$4.78*10^{-6}$
GO:0022402 (BP)	cell cycle process	$1.15*10^{-5}$
GO:0000278 (BP)	mitotic cell cycle	$6.32*10^{-5}$
GO:0007049 (BP)	cell cycle	$7.47*10^{-5}$
GO:0007067 (BP)	mitosis	$2.18*10^{-4}$
GO:0048285 (BP)	organelle fission	$2.18*10^{-4}$
GO:0000280 (BP)	nuclear division	$2.18*10^{-4}$
GO:0000087 (BP)	M phase of mitotic cell cycle	$2.18*10^{-4}$
GO:0005524 (MF)	ATP binding	$6.87*10^{-4}$
GO:0051301 (BP)	cell division	0.0010345834
GO:0032559 (MF)	adenyl ribonucleotide binding	0.0019198435
GO:0030554 (MF)	adenyl nucleotide binding	0.0036673415
GO:0001883 (MF)	purine nucleoside binding	0.0044644562
GO:0001882 (MF)	nucleoside binding	0.0044644562
GO:0000166 (MF)	nucleotide binding	0.0056049633
GO:0022900 (BP)	electron transport chain	0.008686148
GO:0051726 (BP)	regulation of cell cycle	0.009773237
GO:0051276 (BP)	chromosome organization	0.0125079349
GO:0004674 (MF)	protein serine/threonine kinase activity	0.029825514
GO:0003723 (MF)	RNA binding	0.0300973949
GO:0032555 (MF)	purine ribonucleotide binding	0.0308970401
GO:0032553 (MF)	ribonucleotide binding	0.0308970401
GO:0050867 (BP)	positive regulation of cell activation	0.0322112683
GO:0051251 (BP)	positive regulation of lymphocyte activation	0.0322112683
GO:0002696 (BP)	positive regulation of leukocyte activation	0.0322112683
GO:0017076 (MF)	purine nucleotide binding	0.0427998494
GO:0050671 (BP)	positive regulation of lymphocyte proliferation	0.0745291764
GO:0032946 (BP)	positive regulation of mononuclear cell proliferation	0.0745291764
GO:0070665 (BP)	positive regulation of leukocyte proliferation	0.0745291764
GO:0002684 (BP)	positive regulation of immune system process	0.0746925142
GO:0006091 (BP)	generation of precursor metabolites and energy	0.0777034487
GO:0006412 (BP)	translation	0.0793245375
GO:0006605 (BP)	protein targeting	0.0885839432
GO:0044265 (BP)	cellular macromolecule catabolic process	0,0904737852

4. Ensemble feature selection for Treg cell subtype marker identification

This chapter describes the application of ensemble feature selection algorithms to identify Treg cell subtype markers from gene expression data. Our aim is to identify robust surface marker genes based on the meta expression set of Chapter 3 to discriminate between Foxp3$^+$ Treg cell subtypes. Such marker would enable the sorting of specific cell types for further research as these cell types are under suspicion to have different functions in the immune system, and are in principle, potential targets for therapeutic interventions.

The design and implementation of the presented feature selection strategy and the validation using classification was developed in cooperation with immunologists from the German Rheumatism Research Center in Berlin, who designed and performed *wet-lab* experiments. Furthermore, the collaborators contributed with immunological expertise for presenting the biological background, curating the data sets and discussing the results.

4.1. Motivation

For years, Treg cells have been important subjects of medical research due to their crucial role for the adequate regulation of the immune system. Such research would benefit from reliable surface markers allowing the effective discrimination of iTreg and nTreg cells. Being able to perform such discrimination would open the door for detailed functional studies of both cell subtypes and their respective roles within the immune system. Results from such studies could have immediate clinical implications [265, 327]. One prerequisite for clinical applications, like cell transfer therapies, is a clear definition of nTreg and iTreg cells [256, 350].

Up to date, differential expression analysis of Treg specific gene expression data did not reveal stable marker genes to discriminate cell subtypes. There are several studies that have compared the transcriptional profiles of Treg cells to distinguish subsets, but until today only two markers for discriminating nTreg and iTreg cells, *Helios* [101, 298] and *Neuropilin-1* [42, 388] arose. Unfortunately both have been reported to be only partially suitable as markers for nTeg and iTreg cell discrimination [6, 278, 380, 388]. Other strategies to distinct between subtypes were shown to be insufficient, too. For instance *Foxp3*, the master TF of Treg cell cannot function as a surface marker protein, as it is an intracellular marker and requires cell permeabilization for staining [198]. Another so far common strategy is the depletion of CD8$^+$ cells followed by a positive selection of CD25$^+$ cells. These cells are Treg enriched but not a pure fraction [355].

We applied an ensemble of feature selection algorithms to identify yet unknown marker genes that enable the distinction of Treg subtypes. Our approach is adaptable and generalizable to similar problems in related or other fields of cell biology or immunology (see Chapter 2).

4.2. Feature selection

With increasing amounts of data unsupervised machine learning methods became popular such as unsupervised clustering, e.g. k-means [133], SOM [192], unsupervised classification, e.g. SVM [338] or Bayesian reasoning [18]. While unsupervised methods try to infer structures and patterns from unlabeled data, supervised machine learning use labeled data to train a model. One of the most common unsupervised methods is hierarchical clustering, for example to infer structure heatmaps and generate dendograms (see Figure 3.8 as example). Unsupervised methods are used when underlying data structure is unknown and no gold standard or training data is available. Dimension reduction like clustering or principle component analysis (PCA) is a common task for unsupervised methods. Supervised methods use class labels to fit a model based on such labeled data (training set) to separate classes or reduce features necessary to classify unlabeled data (test set).

Here, we aim to identify a small but relevant set of features (genes) for Treg subtype discrimination in the set of expression profiles of thousands of genes while the set of available experiments is limited to a few hundred. Kohavi and John formulated the feature (subset) selection problem as a learning problem of selecting a relevant subset of features, while ignoring the rest [191]. So, the general goal of feature selection is to minimize the set of features to describe a structure [185, 191], which for our problem means to identify single genes or sets of genes whose expression profiles separate two cell sub-populations.

For long, feature subset selection has been a research area within statistics, especially in pattern recognition and machine learning. It is not surprising that feature selection is as much of an issue for machine learning as it is for pattern recognition, as both fields share the common task of classification [130].

Classification aims to infer class labels for unlabeled instances using a generalized model. Identifying a small set of genes that enables such a classification of experiments is goal of our feature selection strategy.

Meanwhile, machine learning was applied in the field of analyzing gene expression data, too. With the establishment of microarray technology, researcher start to apply machine learning techniques to recognize patterns of expression or identify gene groups that describe differences in gene expression.

4.2.1. Methods for feature selection

In contrast to dimension reduction or more general complexity reduction methods, feature selection does not alter the variables and values, but aims to find a subset that best exemplifies the original data [134], while it also aims to offer interpretability to the user.

But, along with the variety of existing methods, related specific secondary objectives exist. Saeys et al. [302] state that the most important ones are:

- avoid overfitting and improve model performance,
- provide fast and effective models,
- gain insides into the underlying processes that generated the data.

Table 4.1.: **Overview of feature selection categories and their major properties.** In column four we list exemplary implementations. The table was inspired and adapted from [143, 302].

	Category	Advantages	Disadvantages	Implementations / algorithms
Filter	Univariate	- Fast and low complexity - No classifier restriction - Computational efficient	- No feature dependencies - No classifier interaction	Information Gain [391] χ^2 [391]
	Multivariate	- Model feature dependencies - No classifier restriction - Still low complexity	- Less scalable to univariate methods - No classifier interaction	Correlation-based feature selection (CFS) [130] Minimum Redundancy Maximum Relevence (mRMR) [275]
Wrapper	Deterministic	- Models feature dependencies - Interaction with classifier - Simple combination of FS and classifier	- Overfitting - Classifier influences final feature set - Problem of local optima (greedy)	Sequential forward selection [166, 187] Sequential backward elimination [166, 187] Gradient-based-leave-one-out gene selection (GLGS) [229]
	Randomized	- Less prone to local optima - Interaction with classifier - Models feature dependencies	- Overfitting - Classifier influences final feature set - Computational intensive	Genetic algorithms [191] Simulated annealing [224] Randomized Hill climbing [391]
Embedded		- Interaction with classifier - Less computational intensive than wrapper methods - Models feature dependencies	- Classifier influences final feature set	Decision trees [191] Weighted Naive Bayes [128] SVM by using the weight vector [124]

Basically, feature selection methods can be divided into the following three categories, (i) filter methods, (ii) wrapper methods and (iii) embedded methods. Table 4.1 summarizes the advantages and disadvantages for the three categories, that we describe in detail in the following subsections.

Filter methods

Filter methods evaluate the relevance of each single feature by regarding the intrinsic properties of the data. Subsequently, features are ranked and those with the lowest relevance are removed [302]. The main difference to the other methods is, that the relevance of single features rather than feature subsets is calculated. Therefore, computation complexity is lower than for methods evaluating subsets. Additionally, the absence of a classifier for evaluating distinction power of the resulting feature set makes the approach suitable for a large feature set, as computation effort does only increase linear. Two subcategories of filter methods exists, namely univariate and multivariate ones. While

univariate methods do not assess relationships or dependencies between features, multivariate methods like Minimal Redundancy Maximal Relevance (MRMR) try to extract minimal feature subsets by reducing the number of identical similar features. As follows we describe four popular filter methods (two univariate and two multivariate):

Information Gain (IG) a univariate method that approximates the conditional probability distribution $P(C|F)$, where C is the class label and F is the feature vector. Thus IG ranks the features regarding their descriptive power to describe the class differences.

Chi^2 statistic (χ^2) evaluates features based on their χ^2 statistics with respect to the class labels [129]. Afterwards, features are ranked, while highest ranked features are potentially best for discriminating the classes.

Minimal Redundancy Maximal Relevance (MRMR) aims to maximize the relevancy for a feature regarding the classes by minimizing the redundancy for each class. As this method considers feature interactions it extracts the best subset of genes describing the classes. The interaction between features is estimated using mutual information (MI) for discrete variable or F-statistics like ANOVA or regression analysis for continuous variables (as present in gene expression data). MI calculates the shared information content of two variables by estimating the amount of information one variable provides about the other. Thus MI is able to asses the redundancy of the subset. For continuous variable the F-statistic estimates the relevance of a gene regarding a class, while correlation measures for gene expression profiles give information about the redundancy of two genes.

Correlation-based feature selection (CFS) aims to find a descriptive set of features by estimating the degree of redundancy among them using correlation. Highly correlated features are removed except for a "representative". In contrast to other methods CFS provides a "heuristic merit" for complete feature subsets. Thus CFS is able to alter the current feature subset based on its current relevance [143].

Wrapper methods

Wrapper methods combine two components into one algorithm, namely a model hypothesis search within the feature subset search. This means that such methods evaluate extracted subsets by dividing data into training and test set and apply classification [130]. Heuristic search algorithms are used to define feature subsets. To this end deterministic and randomized search approaches can been applied. Deterministic approaches offer lower computational complexity, whereas randomization has a lower likelihood of ending up in local optima [143]. As classification is part of wrapper methods, complexity of calculation and the extracted "optimal" feature set relies on the chosen classifier. Examples are:

Sequential forward selection (SFS) also know as Sequential Forward Floating Selection (SFFS), is an incremental method that adds the most descriptive feature to a subset

while removing the least descriptive ones from the complete set. Starting with an empty set, the method consists of roughly three steps. First, the inclusion step, the most descriptive feature among all yet unselected feature is selected. Second, the exclusion step identifies the least significant features among currently selected ones, except for the feature selected in step 1 and removes it. In the third step, again the least significant feature is removed if i) the feature subset consist of at least two features and ii) the feature set evaluation result is worse as of the best other subset so far. If i) and ii) are satisfied a new iteration starts with the first step until no more feature are available.

Sequential backward elimination (SBE) uses the same principle as SFS. In contrast to SFE, SBE is starting with the full set of features and iteratively reduces the set by the least descriptive ones.

Embedded methods

Embedded methods 'learn" the optimal subset by evaluating classification output of feature subsets. In detail such methods aim to derive optimal feature sets using a classifier. In contrast to wrapper methods, the specific classification method (classifier component) is fixed for each embedded method. On one side, this implies that calculated feature subsets highly depend on the applied classifier [143]. On the other side computation is much more efficient than for wrapper methods. Examples are:

Random forest describe a collection of decision trees applied to different data subsets. Results are used to reduce the set of features iteratively by removing those with the lowest average importance in the forest. The smallest forest with the lowest error is finally selected [86].

Support vector machines-recursive feature elimination (SVM-RFE) starts by using all features to separate classes. Gradually all no discriminative features are removed. Discrimination power of a feature is represented by its weight, that is calculated by training an SVM with the current feature set and evaluating its classification performance [143]. Cross-validation is used to avoid local optima caused by data set splitting (into training and test set) [228, 351]

4.2.2. Feature selection for gene marker detection

The application of feature selection methods to gene expression data is a common strategy for several years [302]. The major aim is to distinguish cohorts or groups of experiments in a study using single genes or small groups of genes [32]. Many studies focused on separation of healthy and disease samples, as for examples for cancer [5, 228, 286, 393]. Other fields of application are the detection of drug-targets or disease-diagnosis [386].

Saeys et al. summarized the major categories and feature selection methods in the microarray domain [302]. The authors also stated the advantage of univariate filter methods to be fast and final rankings are intuitive and easy to understand. Another

advantage is, that ranking reflects the discriminative ability of each feature, as this is of high interest when identification of single biomarkers are in the focus of interest.

Ensemble feature selection for increasing robustness

Next to the application of a particular feature selection method, ensembles of methods have been successfully tested to extract and evaluate feature subsets [356]. It has been shown for gene expression and mass spectrometry data, that averaging results from multiple methods increases robustness of results [2, 213, 392]. Furthermore, ensembles can lead to multiple equally well discriminating subsets [393]. Also aggregation of results, or model combination, e.g. boosting, have been tested and shown to improve the robustness and stability of feature selection [302].

Such ensembles of feature selection methods often improve classification results, but increase computation cost. Since large computation resources became available, the impact of computation cost became less important. The subsequent approach combines two filter methods to extract relevant features.

4.3. Feature set evaluation using linear SVM classifier

To evaluate the performance of extracted feature subsets, classification can be used to infer class labels from the exact set of extracted features.

Several classifier methods exist and are suitable to evaluate feature sets, e.g. K-nearest neighbors [277], decision tree or Naïve Bayes [164]. In this work we focus on SVM classifier as these enable robust classification [2]. SVMs are linear classifiers that belong to the general category of kernel methods and are widely used in bioinformatics. Several studies showed their high accuracy on gene expression data (described and reviewed by [24, 136, 338]).

Here, we focused on binary classification, as our aim was to classify experiments from two classes (iTreg, nTreg). The feature vector used for classification was constructed based on the expression values of the genes that were identified as discriminating features. To show the validity of the extracted features we trained a classifier on the extracted feature set and a training set of experiments. Ideally, the classifier is able to infer the correct class of an experiment from a test set using the learned model.

As all feature vectors have the same length, they can easily be projected in the same multidimensional space. Thus, the SVM calculates a hyperplane in this multidimensional space that separates all features of one class from the ones of the other class. Those features, that are used to define the hyperplane from the training data are called support vectors and define the decision boundaries between both classes. The optimal decision is the hyperplane, that maximizes the distances between both classes. An example separation of two classes by a SVM is shown in Figure 4.1.

For details on SVMs see [136, 338], for the specific field of multi-class SVM see [152] and the documentation of LibSVM – on of the most popular libraries implementing SVMs [55].

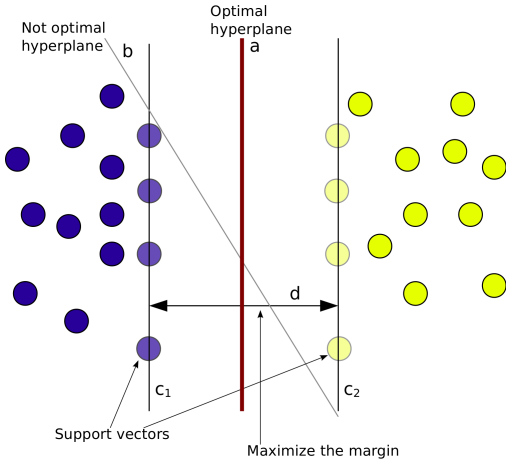


Figure 4.1.: **Support Vector Machines (SVM) aim to find the "best" hyperplane (line a) that separates two classes.** Blue and yellow dots represent entities of two classes, that are separated by a line (a), that maximized decision boundaries (arrow-line d). Two dimension can be separated by a line, like illustrated here, while for higher-dimensional problems a hyperplane separates the classes. A small number of support vectors define the decision function. The support vectors line c_1 and c_2 determine the decision boundaries. Line b is one "not optimal" example of a separating hyperplane.

4.3.1. Classification of imbalanced data sets

A ubiquitous problem of classification is the imbalance of class instances, which means that the majority of instances belongs to one class, where only a small fraction belongs to the other. In such settings most classifier tend to predict the majority class more often then the other [58]. To overcome the *class imbalance problem* solutions on data and algorithmic level were proposed. On the data level it was proposed to sample data, i.e. down-sampling to achieve class balance or oversampling like bootstrapping [232]. On the algorithmic level various approaches have been developed, e.g. class-specific cost variation or instance-weighted support vector machines [58, 371]. A clear advantage of tackling the problem on algorithmic level is that the classifiers keep their general applicability towards imbalanced data sets. This is of importance as for some classification problems the smaller class is typically of higher interest as the other, e.g. in text mining (entity recognition, relationship extraction from text) [357]. In contrast to that, creating test sets/test data with equal class distribution reduces the impact of the imbalance when evaluating extracted feature subsets.

4.4. Detecting robust markers by ensemble feature selection

In this section, we describe our strategy to find a small set of marker genes, that is able to classify unknown cells based on their gene expression profiles. Specifically, we applied

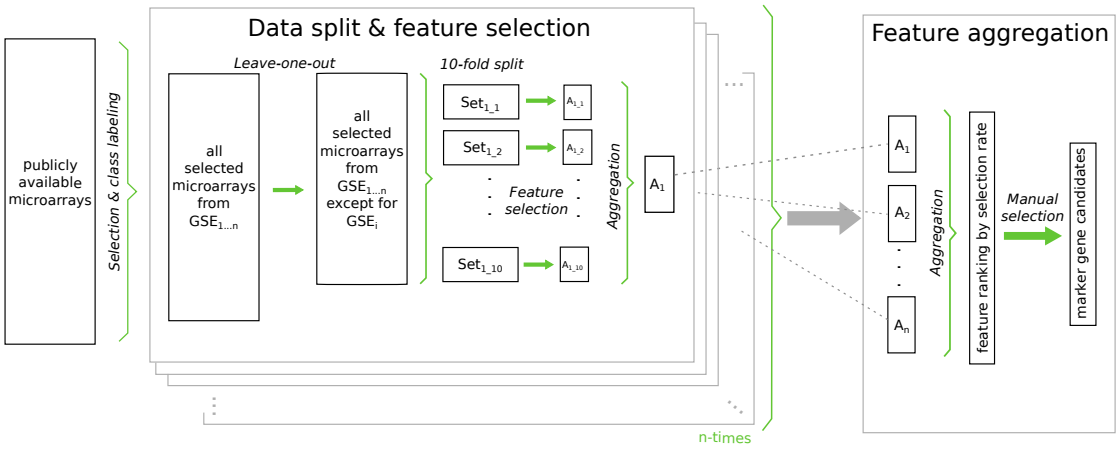


Figure 4.2.: **Data preparation and feature selection pipeline.** Green arrows and braces represent data flows. We performed two data split steps to ensure that all samples from identical studies were not shared between data sets. Feature selection depicts the strongest attributes; here the genes that describe Treg cell subtype class best. GSE is accession number prefix of GEO series, that identifies a specific study.

feature selection to determine marker genes for the distinction between iTreg and nTreg cells based on labeled gene expression data. The meta expression data set from Chapter 3 serves as input. For each experiment in the set, we know whether it is derived from iTreg or nTreg cells.

The basic idea of our approach is to extract those genes, that allow for discrimination between the cell types based on their expression profile. Subsequently, candidates were filtered for those that represent surface marker proteins and therefore enable cell detection by biotechnologies like flow cytometry. The discrimination power of remaining candidates was evaluated using SVM classification. Finally, experimental validation of gene and proteins expressions of the candidates was performed by immunologist using qRT-PCR and Immunoblotting.

The general pipeline of our approach is illustrated in Figure 4.2. The basic steps are:

1. Data preparation & preprocessing, which mainly refers to the meta expression analysis in Chapter 3,
2. Generating overlapping subsets and splitting of these subsets into training and test data for subsequent feature evaluation,
3. Application of feature selection methods and aggregation of results,
4. Evaluation using SVM classification.

The subsequent *wet-lab* validation (Section 4.4.5) is not part of the feature selection pipeline.

4.4.1. Data preparation & preprocessing

Many feature selection methods require discretization of continuous variables [130]. While some methods can handle continuous features, data has to be normalized or transformed to enable comparability between features. As described we used publicly available microarray data from 36 studies. The data was curated and labeled regarding T cell subtype (see Chapter 3, p.34ff).

4.4.2. Data partitioning and splitting into training and test sets

Basic idea of data set partitioning was to generate training and test sets, extracted from independent studies. Thereby, we aim to reduce overfitting of feature selection to our specific data set. The resulting data sets vary in size and balance of nTreg and iTreg experiments.

We performed a two-step splitting of the data set as shown in Figure 4.2 (center part) to achieve two goals:

- increasing robustness of feature selection,
- enabling evaluation of candidate genes using classification on hold-back test data.

First, we performed the data splitting considering study association of the experiments to ensure that feature evaluation is not biased. This was done because expression profiles of experiments (over all genes) from identical studies are often highly correlated or very similar, (e.g. for replicates). As shown in Section 3.4.6 this applies for the presented data set, too. In consequence all experiments from the same studies were assigned to either the training or the corresponding test sets prior the second split. Therefore, we created 36 copies of the meta expression set X. In each copy X'_i all experiments belonging to a certain study i (where $i = 1, ..., 36$) were removed.

Second, each of these 36 subsets was split again, for ten-fold cross-validation. We split each subset X'_i into ten overlapping, but not identical subsets, where each subset included experiments from both classes. For all ten subsets we performed feature selection and aggregated the results to detect those features with highest average discrimination ability and robustness. Overall, we prepared 360 overlapping but not identical subsets for feature selection and 36 independent test sets for latter evaluation using SVMs.

4.4.3. Feature selection methods application and candidate merge

We performed feature selection using an ensemble of two methods, namely Chi^2 and Information Gain (IG). Both are filter methods that calculate the individual weight of a feature independent from other features. As described in Section 4.2.1, the Chi^2 method evaluates features based on their χ^2 statistics with respect to the class labels [129]. The IG based approach uses as selection criterion the information gain for all pairs of genes [391].

We used the implementations provided by the Weka 3: Data Mining Software (Weka) [129] with default settings. Both methods were applied separately to all 360 subsets.

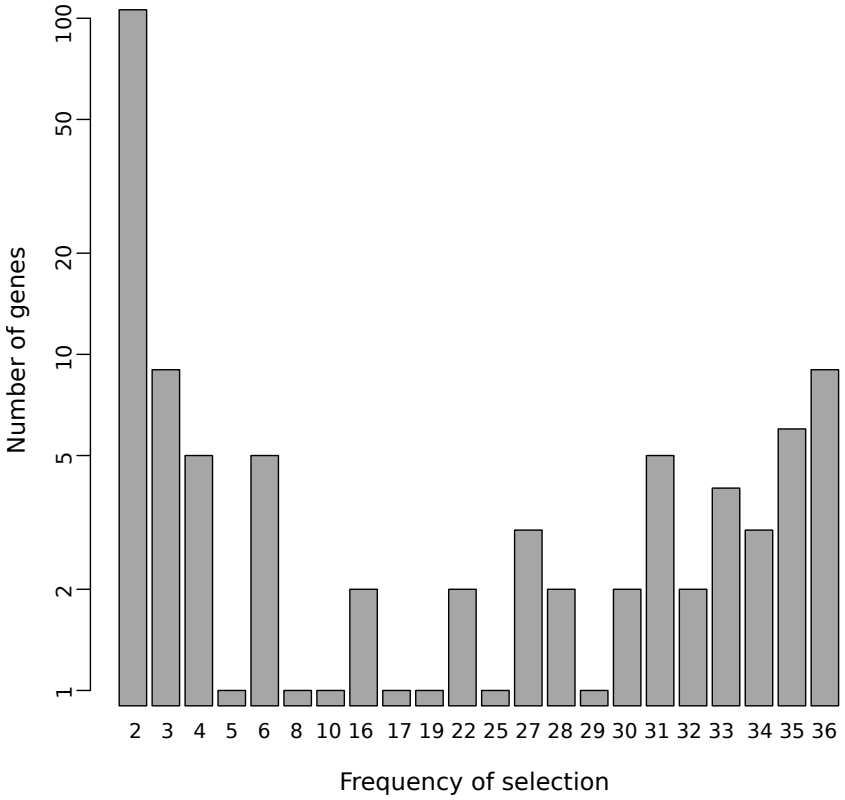


Figure 4.3.: **Frequency of simultaneously selected features (genes) by both methods on different subsets.** We extracted those genes as candidates that were selected as features in more than 50% (here: 18 of 36) of the subset analyses. This cut-off led to 41 candidate genes.

Each method produced a list of genes ranked by discriminative power, which we cut off after 100 genes. This gave rise to 36 times 10 sets of 100 genes for each method.

For extraction of candidates, the calculated gene sets were combined as follows. First, for each of the 36 subsets the intersection of both feature selection methods was calculated for each of the 10 sets. Next, genes were ranked by the number of occurrence in these 10 sets, and only those were kept that occurred in the majority of the selected sets, i.e. six times. Combining the resulting 36 gene sets, we considered those candidates that are present in at least 50% of the subset evaluations (i.e. 18 times, as shown in Figure 4.3).

Table 4.2 shows the list of the 41 most top-ranked genes from feature selection and aggregation. To enable the application as cell type markers we investigated their biological characteristics. Central point of additional manual analysis was to check whether the candidates represent cell surface markers, as only these could be utilized for cell sorting. We identified six genes that were detected to have strong discriminating power and code for surface proteins, namely *Cd79b, Cd97, Ctla-2b, Emp1, Gabrr2* and *Tnfrsf13b*.

Table 4.2.: **Extracted marker gene candidates** All marker gene candidates identified by feature selection step of analysis pipeline. Bold printed candidate genes were manually selected and evaluated regarding cell type marker functionality, here categorized as surface molecules. Columns "pValue" and "fc(iTreg - nTreg)" contain information from t-test analysis and indicate whether genes are differentially expressed between cell types (double asterix). Table from related publication [198].

Ensembl Gene ID	MGI symbol	MGI Description	pValue	fc(iTreg - nTreg)		Category
ENSMUSG00000040592	**Cd79b**	**CD79B antigen**	**4,47E-005**	**-1,009**	**	**Receptors**
ENSMUSG00000002885	**Cd97**	**CD97 antigen**	**1,37E-006**	**-1,117**	**	**Receptors**
ENSMUSG00000074874	**Ctla2b**	**cytotoxic T lymphocyte-associated protein 2 beta**	**2,38E-010**	**2,667**	**	**Receptors**
ENSMUSG00000030208	**Emp1**	**epithelial membrane protein 1**	**5,20E-011**	**2,537**	**	**Surface molecules**
ENSMUSG00000023267	**Gabrr2**	**gamma-aminobutyric acid (GABA) C receptor, subunit rho 2**	**9,38E-007**	**-1,155**	**	**Receptors**
ENSMUSG00000010142	**Tnfrsf13b**	**tumor necrosis factor receptor superfamily, member 13b**	**1,67E-002**	**0,564**		**Receptors**
ENSMUSG00000029687	Ezh2	enhancer of zeste homolog 2 (Drosophila)	9,04E-005	0,817		Transcriptional repression
ENSMUSG00000030528	Blm	Bloom syndrome, RecQ helicase-like	1,19E-008	1,304	**	DNA replication
ENSMUSG00000000028	Cdc45	cell division cycle 45	1,40E-008	1,041	**	DNA replication
ENSMUSG00000031585	Gtf2e2	general transcription factor II E, polypeptide 2 (beta subunit)	2,30E-008	1,505	**	DNA replication
ENSMUSG00000033910	Gucy1a3	guanylate cyclase 1, soluble, alpha 3	6,33E-006	-1,132	**	DNA replication
ENSMUSG00000017716	Birc5	baculoviral IAP repeat-containing 5	2,86E-010	2,138	**	Apoptosis
ENSMUSG00000025962	Fastkd2	FAST kinase domains 2	3,27E-002	-0,672		Apoptosis
ENSMUSG00000027496	Aurka	aurora kinase A	1,83E-009	1,321	**	Cell cycle
ENSMUSG00000042111	Ccdc115	coiled-coil domain containing 115	1,33E-007	0,903		Cell cycle
ENSMUSG00000023067	Cdkn1a	cyclin-dependent kinase inhibitor 1A (P21)	6,17E-012	2,5	**	Cell cycle
ENSMUSG00000037725	Ckap2	cytoskeleton associated protein 2	1,65E-011	1,574	**	Cell cycle
ENSMUSG00000020649	Rrm2	ribonucleotide reductase M2	5,52E-010	2,394	**	Cell cycle
ENSMUSG00000022322	Shcbp1	Shc SH2-domain binding protein 1	5,58E-012	2,837	**	Cell cycle
ENSMUSG00000024892	Pcx	pyruvate carboxylase	1,15E-005	0,715		Cell Metabolism
ENSMUSG00000062070	Pgk1	phosphoglycerate kinase 1	4,63E-011	1,405	**	Cell Metabolism
ENSMUSG00000027999	Pla2g12a	phospholipase A2, group XIIA	7,21E-003	0,705		Cell Metabolism
ENSMUSG00000020089	Ppa1	pyrophosphatase (inorganic) 1	3,49E-003	0,666		Cell Metabolism
ENSMUSG00000024378	Stard4	StAR-related lipid transfer (START) domain containing 4	4,22E-007	1,056	**	Cell Metabolism
ENSMUSG00000023087	Ccrn4l	CCR4 carbon catabolite repression 4-like (S. cerevisiae)	8,02E-007	1,118	**	Cicardian clock
ENSMUSG00000021575	Ahrr	aryl-hydrocarbon receptor repressor	2,01E-007	1,013	**	Drug metabolism
ENSMUSG00000038776	Ephx1	epoxide hydrolase 1, microsomal	4,25E-011	-1,596	**	Drug metabolism
ENSMUSG00000024480	Ap3s1	adaptor-related protein complex 3, sigma 1 subunit	1,25E-006	1,936	**	Golgi
ENSMUSG00000029076	Sdf4	stromal cell derived factor 4	7,56E-004	-1,394	**	Golgi
ENSMUSG00000024143	Rhoq	ras homolog gene family, member Q	4,53E-007	2,026	**	GTPase
ENSMUSG00000027379	Bub1	budding uninhibited by benzimidazoles 1 homolog (S. cerevisiae)	7,43E-012	2,401	**	Mitosis
ENSMUSG00000040084	Bub1b	budding uninhibited by benzimidazoles 1 homolog, beta (S. cerevisiae)	4,36E-008	1,385	**	Mitosis
ENSMUSG00000029910	Mad2l1	MAD2 mitotic arrest deficient-like 1	2,67E-009	1,398		Mitosis
ENSMUSG00000015880	Ncapg	non-SMC condensin I complex, subunit G	1,17E-012	1,749	**	Mitosis
ENSMUSG00000038943	Prc1	protein regulator of cytokinesis 1	2,02E-011	1,989	**	Mitosis
ENSMUSG00000028718	Stil	Scl/Tal1 interrupting locus	3,77E-012	2,13	**	Mitosis
ENSMUSG00000020721	Helz	helicase with zinc finger domain	3,93E-004	-1,499	**	RNA metabolism
ENSMUSG00000028970	Abcb1b	ATP-binding cassette, sub-family B (MDR/TAP), member 1B	3,01E-001	0,312		Drug metabolism
ENSMUSG00000029463	Fam216a	family with sequence similarity 216, member A	3,20E-007	0,786		
ENSMUSG00000029720	Gm20605	predicted gene 20605	2,37E-008	-0,947		
ENSMUSG00000075271	Ttc30a1	tetratricopeptide repeat domain 30A1	5,72E-008	1,025	**	

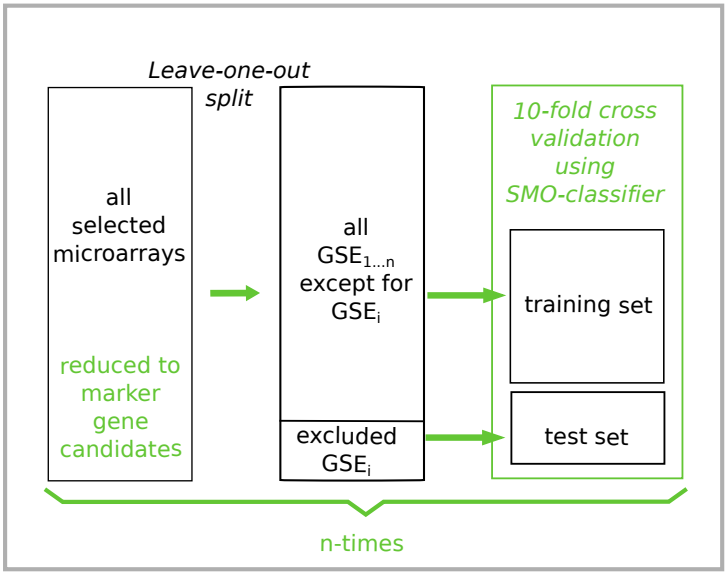


Figure 4.4.: **First phase of evaluation: Classification of experiments based on potential marker genes.** The complete meta expression set was used for leave-one-out cross validation, but samples from the same study are not shared between training and test sets.

4.4.4. Subset evaluation using SVM classifier

The set of candidate genes were further validated computationally. As we used two filter methods for feature selection, we now aim to evaluate the discrimination power of the extracted candidates using a SVM classifier. For computational evaluation, we used all six candidates as features for a support vector machine classifier [73] to discern iTreg from nTreg cell samples. Evaluation was performed in a leave-one-out manner (Figure 4.4), and average results across all folds are reported. As SVM implementation we used the sequential minimal optimization algorithm (SMO) [279] implemented in Weka.

In Table 4.3 we report average results for all metrics across all folds following the described evaluation in Figure 4.4. We evaluated using metrics, precision, recall, sensitivity, specificity, accuracy and F-measure. The metrics for binary classification as it is used here, are described as follows:

True positive (TP) depicts the number of correctly predicted class 1 labels.
False positive (FP) depicts the number of falsely predicted class 1 labels.
True negative (TN) depicts the number of correctly predicted class 2 labels
False negative (FN) depicts the number falsely class 2 predicted labels
Precision is the proportion of correctly predicted class 1 labels to all existing class 1 labels.

$$Precision = \frac{TP}{TP + FP}$$

Sensitivity / Recall is the proportion of correctly predicted class 1 labels among all

Table 4.3.: **Performance metrics of validation using SMO classification results for each cell types over all folds.** Metrics were calculated for both candidate sets, six selected marker genes and full set of 41 candidates. Specificity and sensitivity for both classes are "diagonal identical", as a correctly positive inferred instance of one class (*true positive*) is a correctly negative inferred instance (*true negative*) in the other class.

	6 candidate genes		41 feature selection genes	
	iTreg	nTreg	iTreg	nTreg
Precision	0.8750	0.9746	0.8571	0.9504
Sensitivity/Recall	0.8750	0.9746	0.75	0.9746
Specificity	0.9746	0.8750	0.9746	0.75
Accuracy	0.9577	0.9577	0.9366	0.9366
F-measure	0.8750	0.9746	0.8	0.9055

predicted class 1 labels.

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity is the proportion of correct predicted class 2 labels among all class 2 labels.

$$Specificity = \frac{TN}{TN + FP}$$

Accuracy is the proportion of the sum of correctly class 1 predicted inferred interactions (true positives) plus correctly predicted class 2 labels.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

F-measure is the harmonic mean of precision and recall.

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Furthermore, weighted metrics are visualized in Figure 4.5. In contrast to Table 4.3, metrics shown in Figure 4.5 were calculated regarding the specific class balance of the test sets to illustrate variance of measures between test sets. Additionally, we evaluated the performance of the complete extracted set of 41 candidates to show validity of the feature selection.

4.4.5. Experimental validation of detected Treg subtype marker gene candidates

For *wet-lab* validation of extracted subtype marker, murine Treg cells were used, because in silico analysis was also carried out with mice transcriptome data. Also, performing the

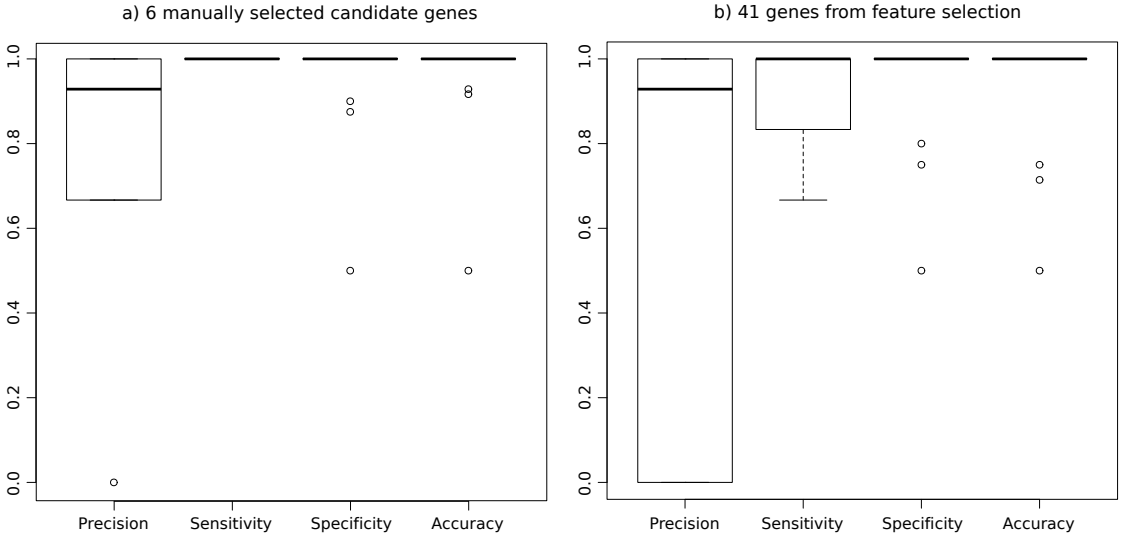


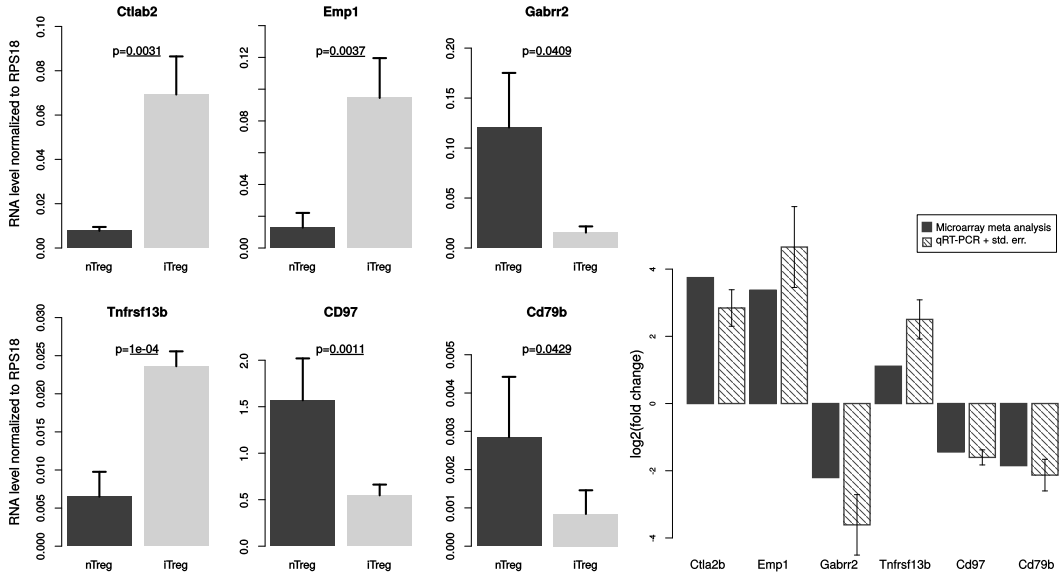
Figure 4.5.: **Weighted performance metrics of SVM-based SMO classification.** Leave-one-out cross validation was performed. Features were restricted to **a)** 6 manually selected genes out of **b)** 41 genes as found by feature selection. For each test set the weighted mean was used to balance metrics derived for both Treg subtype classes. In contrast to Table 4.3, metrics were calculated for each test set classification separately to illustrate variances between the sets.

validation in mice has the advantages of inbreed consistency and the far greater number of available data sets, which is very important for our strategy to increase robustness of predictions by grounding in a large set of heterogeneous experiments. However, we are aware that direct transfer of our results to humans is not possible, because not all genes in mice have a homologous gene in human.

Validation of data set derived expression differences between cells types on RNA level

To verify gene expression intensities of the six candidate genes in an independent experiment via qRT-PCR nTreg cells were sorted from Foxp3-IRES-EGFP mice and iTreg cells were generated *in vitro* by cultivating naïve T cells with IL-2 and TGF-β. There is a remarkable consistency of the results conclusions from the microarray analysis (Figure 4.6). qRT-PCR indicates even higher significance of changes between iTreg and nTreg cells for the selected genes than our microarray analysis (Figure 4.6b). The genes *Ctla2b*, *Emp1* and *Tnfrsf13b* had a higher expression in iTreg cells compared to nTreg cells, whereas *Gabrr2*, *Cd97* and *Cd79b* had a higher expression level in nTreg (Figure 4.6a). Thus, successful validation using qRT-PCR shows that extracted gene expression profiles of the candidates enables the discrimination between the Treg cell subtypes, if cells are treated similarly to those in the microarray experiments used for feature

selection.



(a) Relative mRNA expressions (b) Fold changes

Figure 4.6.: **a) Relative mRNA expression of candidate genes was analyzed by qRT-PCR.** mRNA was extracted from cultured iTreg and *ex vivo* nTreg cells of Foxp3-IRES-GFP mice and normalized to RPS-18 expression (n=6). The results show high consistency for predicted markers between meta-analysis and validation experiments; validation and original experiments were performed under similar conditions (non-cultivated nTreg cells). **b) Comparison of fold changes for candidates derived from meta-analysis and qRT-PCR experiments.** Rough-hatched bars show log_2 fold changes calculated from meta-analysis. Fine-hatched bars depict fold changes calculated from qRT-PCR between cultured iTreg and *ex vivo* nTreg cell.

Validation of candidates on protein level using Immunoblotting

To analyze protein expression of the candidate genes, we performed immunoblot experiments with available antibodies. Unfortunately, only for Cd97 and CD79b working antibodies were available. EMP1 showed a slightly increased expression in nTreg cells compared to iTreg cells (Figure 4.7), whereas the receptor GABRR2 showed a clear difference in protein expression between both subtypes. The results show that protein expression levels obtained from immunoblot analysis of EMP1 and GABRR2 are able to distinct *in vitro* Treg and *ex vivo* nTreg cells as predicted by own *in silico* analysis.

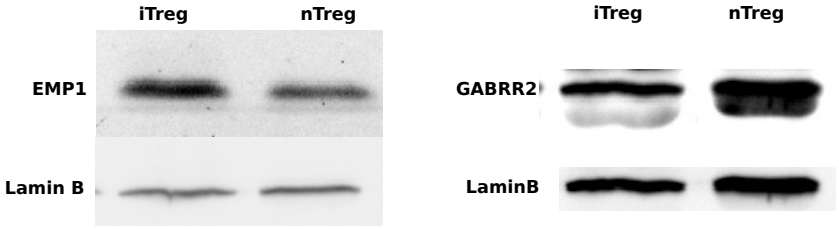


Figure 4.7.: **Immunoblot analysis of EMP1 and GABRR2.** Analysis revealed stability of differences in protein expression between cell subtypes for both candidates. Expressions of both proteins where subtracted by background signal and normalized to LaminB expression.

Implication from *wet-lab* validations

We identified six candidate genes encoding surface proteins, namely *Cd79b, Cd97, Ctla-2b, Emp1, Gabrr2, Tnfrsf13b*. For these genes we confirmed the expression profiles from microarray meta-analysis via qRT-PCR (Figure 4.6), while using comparable conditions as in the baseline microarray experiments.

Unfortunately, it is not easy to infer applicability of the markers for discrimination of these cell types as the conditions which where used by the majority of the experiments in the meta data set do not or poorly reflect nature. A major problem is, that the iTreg cell were induced and harvested under certain cultivation conditions which were not applied to the *ex vivo* nTreg cells. Thus, validation experiments did indeed validate the presented machine learning strategy to find differences between cell types and also showed the validity of the extracted marker candidates.

Further *wet-lab* experiments revealed that observed differences are likely to be originated from cultivation differences between cells (data not shown here). Nevertheless, it was shown (see Figure 4.8), that three of the extracted marker (*Cd79b, Ctla-2b, Gabrr2*) have potential marker quality. Because they were also differentially expressed when experiments were performed under identical conditions (with additionally cultured *ex vivo* nTreg and *in vitro* iTreg).

4.4.6. Discussion

In contrasts to differential expression analysis the presented feature selection approach of marker gene identification is not based on fold change height, rather than ability to discriminate between cell types. In fact, the chosen strategy of using top ranked genes from ensemble feature selection uses an arbitrarily set up threshold, which is a common issue on ranking methods [86]. But in contrast to differential expression analysis using arbitrarily chosen minimal fold changes, the threshold is set after assessing genes specific marker quality.

Another feature selection based strategy is to reduce the gene set in advance by those genes that do not code for surface marker. That refers to the first point of the "feature selection checklist" by Guyon et al. [124].

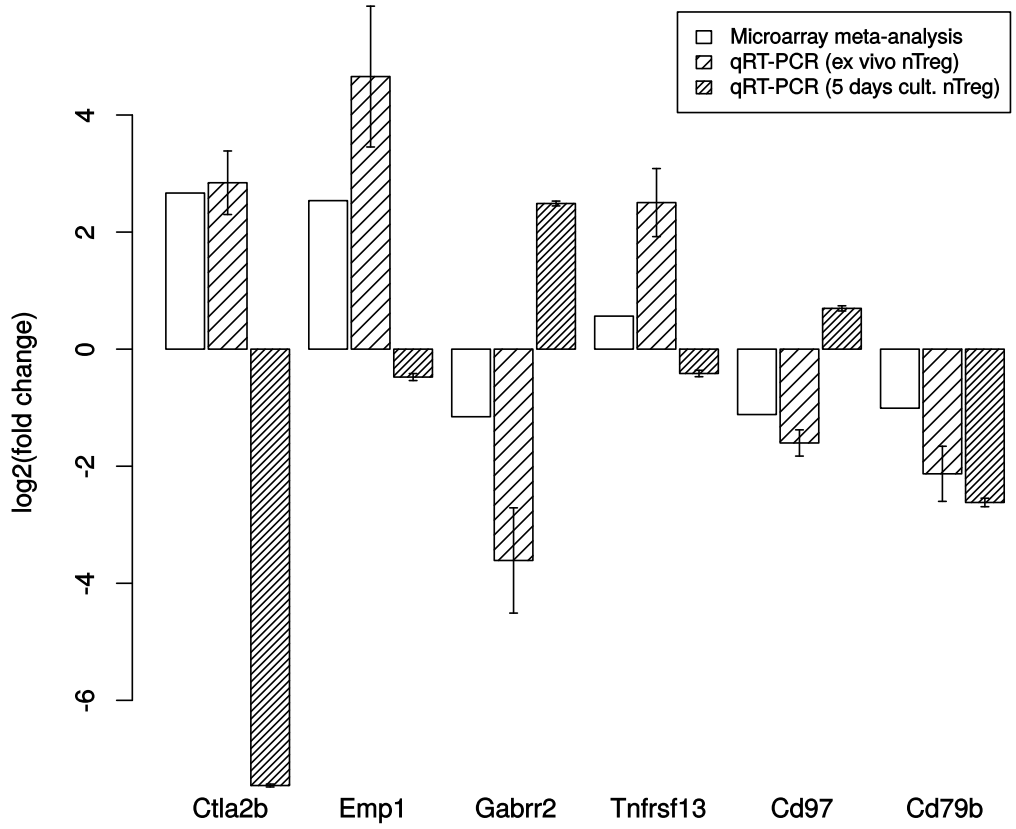


Figure 4.8.: **Gene expression log_2 fold changes including cultured *ex vivo* nTreg cells.** Non-hatched bars show fold changes calculated from Bioinformatics meta-analysis. Rough-hatched bars depict fold changes calculated from qRT-PCR between iTreg and not cultured nTreg cell. Fine-hatched bars depict fold changes from qRT-PCR between iTreg and five days cultured nTreg cells.

But first, as we did not obtain a gold standard for all surface marker proteins, such set reduction could have led to exclusion of marker genes by false annotation. Second, next to identifying marker candidates, other, more general goal of this work was to shed more light on the differences between both cell subtypes, as this includes more differences as the expression of certain genes. The full list of retrieved features could also be used for further investigation.

Another drawback is information leakage, i.e. that training and test set data for validation was extracted from the same set of experiments (shown in Chapter 3). Nevertheless excluding complete experiments from feature selection would have substantially reduce the amount of input data and thus the robustness of our methods. Figure 4.9 shows the size of the included experiments to illustrate the frequency of small experiments. Excluding small studies would have resulted in inefficient training and test set size for classification, while excluding large experiments would have effected feature learning.

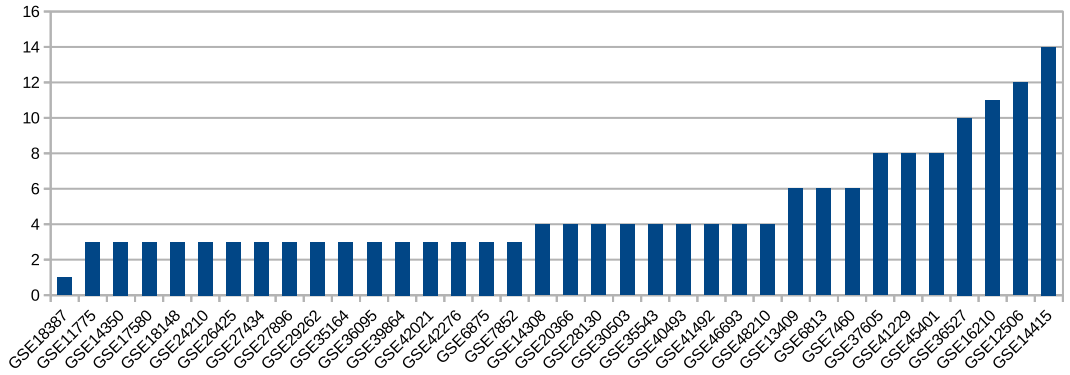


Figure 4.9.: **Number of microarray experiments per study.** The majority (72%) of integrated studies/ study parts are rather small, containing less than five microarrays.

4.5. Summary

We reported on a machine learning approach to identify potential surface markers for separating *in vitro* differentiated iTreg cells from *ex vivo* nTreg cells [198]. Different feature selection methods were analyzed and compared. We applied two methods to the previously described meta-expression set (in Chapter 3). We hypothesized that markers which are statistically discriminative under a wide variety of experimental conditions should be more robust than those from a single experiment only. Using statistical learning algorithms, we identified 41 candidate genes, which are highly discriminative between nTreg and iTreg cells. Among these 41, six genes encode for surface proteins and could thus be potentially applied as sorting markers in cell. To verify our machine learning results we first validated these candidates using classification by SVMs and secondly performed *wet-lab* experiments (qRT-PCR on mRNA level and Western Blotting on protein level) with murine *in vitro* generated iTreg and *ex vivo* isolated nTreg cells. However, since the vast majority of the nTreg data was generated using uncultivated nTreg cells, observed differences could be an artifact of the experimental procedures and not of the underlying biology.

Nevertheless, our approach successfully identified genes for discrimination between cell types. Furthermore the approach is easily extendable to other biological context and thus offers an alternative to existing methods like differential expression analysis.

5. Reconstruction of gene regulatory networks

In this chapter we focus on methods for reconstruction of gene regulatory networks (GRNs). As described in Chapter 2, current knowledge on gene regulation of Treg cells is insufficient to explain all regulatory effects. While various methods for reconstructing GRNs were published in the last decade, their performance varies strongly between different data sets [137, 246]. Since we are interested in Treg cell specific gene regulatory interactions, we aim to evaluate the performance of different reconstruction methods applied to a Treg cell specific dataset. Knowledge on T cell subtype specific GRNs may elucidate the differences and can lead to a novel approach for understanding and controlling the differentiation (described in Chapter 2).

The chapter is organized as follows: First we introduce the general concept of GRNs. Second, we provide an overview of available methods and tools for the reconstruction based on gene expression data. Third, we infer networks from steady state expression data by applying a subset of the presented tools to the Treg cell specific meta-expression data introduced in Chapter 3. Finally, we evaluate and compare tool performances using two public interaction databases and examine the most commonly inferred interactions.

5.1. Motivation

Research on networks and especially network visualization has a strong history in biology [7]. Since, the knowledge about genes, proteins and their interactions grew rapidly over the last decades, it has become imperative to find proper forms of presentation. As in other fields, networks aim to visualize the coherence of biological entities. When the networks became larger and more dense, network analysis was needed and adopted from other fields, e.g. social network analysis (SNA) [376]. In these early days, the applied methods and metrics were topology-based [21, 87, 94, 169, 288]. Later, quantitative measures like correlation and co-expression were focused. Over the last 15 years, gene co-expression networks, protein-protein interaction networks and cell-cell interaction networks became the most important types of biological networks [402].

Today, GRNs are an important type of biological networks as they aim to explain the regulatory mechanisms (edges) between genes (nodes). The term and concept gene regulation is under permanent progression and actors and mechanisms are continuously added. Understanding regulatory mechanisms on large scale is one of the most important goals in systems biology. Gene regulation describes the mechanisms influencing gene expression, i.e. binding, transcription and translation.

The translation from genes to proteins is based on gene regulatory mechanisms and essentially influenced by specific elements. Via DNA binding, cis-regulatory elements like TFs control the transcriptional events of target genes. Such elements can bind directly to the DNA or indirectly influence transcription via co-factorial binding or as part of compounds [242] and thus influence the expression of genes. The identification of such regulatory mechanisms is the aim of GRN reconstruction. GRNs enable the flexible representation of gene regulation. They are extendable by additional concepts (edges and nodes) like different types of edges representing different kinds regulatory influence or linking nodes, beside genes, to integrate different biological levels, e.g. miRNAs or proteins.

In this work, the terms reconstruction and inference describe the process of uncovering direct and indirect interactions between regulatory elements, such as activation, inhibition or binding. The edges between genes in GRNs are often termed gene regulation, regulatory event, regulatory dependency or more general interaction or relation. Here, we focus on the transcriptome level, where gene expression analysis is used to detect state changes of genes or can be used to calculate correlations between gene expressions. Although, GRN reconstruction is not a new topic in the field of computational biology, still a number of problems remain unsolved until today:

- The quality and performance of GRN reconstruction depends on the amount of available input data. Producing large data sets for higher organism, like human or mouse experiments are time, and money consuming.

- The "small m large p problem", still accounts and requires pre-selection and reduction of the target gene set prior to the actual reconstruction. This rule says, that the number of queried regulators (e.g. genes) is limited by the number of input experiments (m) used to determine regulatory effects [30]. To overcome the problem heuristics are required or the set of potential target genes must be reduced, as it is the case for ODE-based approaches [31].

- The evaluation of inferred GRNs is an open problem, as each newly inferred regulatory relation must be experimentally validated. High fractions of incorrectly inferred regulatory relations increase cost and reduce efficiency. No gold standards for evaluation of networks of higher organisms like mammals exist and reliable estimations about size and density of the true networks are unavailable. Simulated data sets and validated data sets are used for comparative studies and evaluation of GRN reconstruction methods to overcome this shortage. A popular platform providing such data is the DREAM challenge - an annual, public competition that focuses on comparing submitted algorithms and tools based on predefined question and provided data [247].

Different methods, approaches and implementation have been published over the years and since the first DREAM challenge - annual challenges of biological interaction and network inference called for submissions in 2006, more than 70 paper were published[1].

[1]List of publications, available here: `http://dreamchallenges.org/publications/`

During the first years of the challenge (up to DREAM4 in 2009), biological data was mainly simulated (*in silico*) and only partially "real" experimental expression data from simple organism like E.coli [246] was applied.

Since this time, various authors published algorithms and tools for GRN reconstruction. In the following section we describe different methods and compare their basic characteristics (Table 5.4, p.80)

5.2. Methods for GRN reconstruction

In this section, different methods of network inference will be presented. The methods are explained regarding their application towards biological data and the reconstruction of gene regulatory networks.

GRNs methods can be categorized into those who deal with steady state data and those with handle time-series data, often referred as dynamic data. Another important distinction is the ability of handling priors, e.g. known interactions. Such priors are used by many methods to set up an initial network structure as baseline for subsequent reconstruction.

5.2.1. Formal definition of GRNs

A GRN is a graph $G = (V, E)$, of nodes V, that represent genes, and edges E, that represent regulatory interactions between genes. An edge $e \in E$ is defined by two adjacent vertices $v_1, v_2 \in V$.

GRNs can either be undirected or directed (Figure 5.1) to indicate the actor of a regulation and its target gene. TFs or other cis-regulatory elements are connected to their target genes, while they can be targets of other regulators, simultaneously. This includes them-self in case of loops. Loops are frequent structures in GRNs for example the IL2 cytokine loop during T cell differentiation [44].

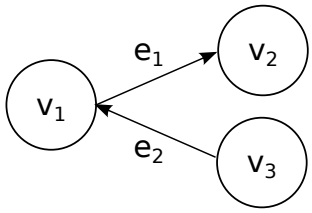


Figure 5.1.: **Directed GRN with three nodes v_1, v_2, v_3 and two edges e_1, e_2.** Node v_1 is the source of $e_1$and target of e_2, v_2 is targeted by a regulatory event initiated by v_1. v_1 might represent a TF that is target of another TF, here v_3. Regulatory strength can be represented as edge weights assigned to e_1 and e_2.

GRN can also be weighted graphs to provide quantitative measure to a regulatory interaction. Weighted graphs are defined as: $G = (E, V, w)$ where $w : E \to W$ is a mapping function of edges to discrete or continuous values (weights).

Table 5.1.: **Comparison of steady state data and time-series data**

Steady-state data	**Time-series data**
• Data measured at a fixed time point or process state	• Collection of experiments/ measurements over time
• Used to compare states, e.g. drug-response, different cell-type, tissue samples	• Adds a temporal dimension of signal capturing,e.g. to cover different temporal states during a biological process

Steady-state and time-series data

In this study, we focused on GRN reconstruction from steady state data, as large time-series data for T cells (especially Treg cells) is not publicly available. However, it is important to mention, that using time series data is more straight forward than steady state data [315], because it allows for capturing the dynamics of gene regulatory processes and is able to reveal directional interactions to indicate the cause-and-effect relationships among genes [373]. For a short comparison of GRN reconstruction based on steady state and time-series data see Table 5.1.

5.2.2. Co-expression or correlation networks

In 1998, Eisen et al. described co-expression as similarity in pattern of gene expression [92]. The authors used hierarchical clustering to structure expression pattern of genes and visualized co-expression using heatmaps (e.g. Figure 3.8), a still often used representation of expression profiles, until today. A few years later, researchers started to extract GRNs from gene expression data by detecting co-expression and visualize the correlation between genes as networks [52, 90, 341]. Gene expression studies were typically performed to compare the expression level of genes between healthy and diseased tissue [81]. Later it became also popular to compare gene expression levels at different time points. However, the basic idea is to test whether genes share common expression profiles over different measurements (e.g. states or time-points). The similarity of gene expression profiles is assessed by testing for significance of expression changes between state. Major challenge on inferring co-expression networks is to define a threshold as lower bound for assuming two genes as biological meaningful connected, i.e. the translation into the binary relation “connected/unconnected” [402]. A common approach to define the set of connected genes is calculating the pairwise correlation between all genes and using a correlation threshold (above 0.5) to decide whether two genes are connected or not. If no threshold is defined the network is fully connected, i.e. all genes are connected to all other genes.

Weighted co-expression networks

Weighted co-expression networks are an extension of the general co-expression network concept by edge weights. Thereby, the value representing the relevance of the co-expression is assigned to each edge in the network. Already in 2000, Butte and Kohane suggested to use the significance of the correlation as edge weight [46]. They stated that the binary correlation values do not give any information about the biological meaning of the interaction which hampers the comparison between different studies. In contrast to that, values depicting the significance of a co-expression allow for inter-study comparison [52]. Zhang and Horvath described different approaches to construct co-expression networks [402]. First, unweighted co-expression networks, wherein hard-thresholding is used to remove edges not exceeding the minimal co-expression criteria. Such networks can be represented as a binary adjacency matrix. Second, weighted networks, that are based on soft-thresholding, i.e. all edges with values above a certain minimum are included and the corresponding co-expression measure is assigned as edge weight. It was also suggested to use only those function parameters for soft-thresholding, that lead to scale-free network topologies. They justified this by several studies, that revealed that the majority of various kinds of biological networks (e.g. pathway or metabolic networks) are scale-free.

Measures of co-expression

A common method to quantify co-expression between entities is to apply by correlation measures between their expression profiles. Correlation metrics like Pearson Correlation Coefficient (PCC) or Spearman's rank Correlation Coefficient (SCC) are calculated for each pair of genes. Additionally, estimation of statistical significance of correlation or simple sorting of correlation values can be used to reduce the networks by removing edges representing low correlations.

The PCC is a linear correlation coefficient that returns a correlation measure p between -1 and 1 , where -1 represent full anti-correlation, 1 identity, and 0 no correlation. PCC uses the covariance (cov) of the gene expressions (x, y) to quantify the strength of co-expression:

$$p_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}, \text{ where } \sigma_x \text{ is the standard deviation of } x. \tag{5.1}$$

The SCC describes the correlation between two variables as monotonic function.

SCC is defined as the PCC for ranked variables, as this means:

$$r_S = p_{r_x,r_y} = \frac{cov(r_x,r_y)}{\sigma_{r_x}\sigma_{r_y}} \tag{5.2}$$

where r_x and r_y are the ranks of x and y. All SCC values r_S are $-1 \leq r_S \leq 1$, where -1 denotes anti-correlated, 1 full correlated and 0 non-correlated entities, as for the PCC. Table 5.2 shows criteria for choosing either PCC or SCC as correlation measure.

Table 5.2.: **Properties of correlation metrics, PCC and SCC [95].**

Pearson Correlation Coefficient	**Spearman's rank Correlation Coefficient**
• Continuous values	• Interval or ratio level or ordinal values
• Entities are linearly related	• No assumptions on distribution
• Values are bivariate normally distributed	• Based on the ranked values rather raw data

Construction of co-expression networks

Co-expression networks are constructed based on correlation metrics. Correlation measures are calculated for all pairs of genes based on methods such as the presented Pearson or Spearman correlation measures. Calculating all pairwise correlations results in a symmetric adjacency matrix of all genes (nodes). The unique (upper or lower) triangle of the matrix represents a network with n nodes and $e = \frac{n^2}{n-1}$ edges.

Different approaches have been developed to reduce the amount of edges. For example,

- remove edges with a correlation values below a (prior defined) threshold,
- include only edges that represent statistically significant correlation values (lower bound for p-values),
- forbid triangles (three fully connected nodes), by removing the weakest edge,
- use additional criteria independent of correlation.

5.2.3. Bayesian networks

Another common approach to estimate the regulatory relationships between genes is learning Bayesian network (BN) from steady state high-throughput expression data [105]. In general, BNs model the qualitative properties of GRNs by using a combination of probability estimation and graph theory [53]. The objective of network inference using BNs is to find the network that "best describes" the probability distribution over the set of variables [104]. Thereby, each gene is treated as a random variable and the network graph represents the joint probability distribution over the set of variables.

Definition

A BN represents the informational or causal dependencies among the variable (nodes) of the network (see Figure 5.2). The dependencies are quantified by conditional probabilities of each node given its parents in the network. A BN is defined as

$$BN = (G, P), \tag{5.3}$$

where $G = (V, E)$ is a directed acyclic graph (DAG) and P is a set probability distribution. Each node V_i in the BN is described by the entirety of its concrete set of ancestors Π_{V_i}, represented as a set of random variables. For GRNs, this means that the expression

profile of gene v_1 is explained by its incident genes $\Pi_{V_1} = v_{1_1}...v_{1_n}$, so called parents. [263]

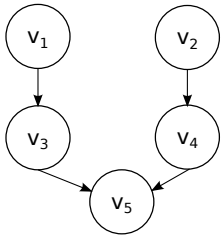


Figure 5.2.: **A Bayesian network representing conditional probabilities among variables.** The graph illustrates a simple BN, where the $v_1...v_5$ represent genes and the causal relationship amongst them. In this example the expression of v_3 and v_4 depends on v_1 and v_2 respectively, where the expression of v_5 depends on the common expression of v_3 and v_4. Despite no direct relation between v_1 and v_5, it is clear, that v_5 depends on the expression of v_1 and this relation is mediated by v_3.

This is known as the Markov property defining the assumption, that it is sufficient to know the current state of a stochastic process (and not its past history) to predict its futures state [95]. This enables the representation of the joint probabilities distribution $P(G)$ of the random variables as a product of conditional probability distributions:

$$P(G) = \prod_{i=1}^{p} P_{G_i}(G_i|\Pi_{G_i})), \tag{5.4}$$

where Π_{G_i} is the product of the set of parents of i [273]. BNs have three fundamental structures, *converging connection, serial connection* and *diverging connection* to represent all possible configurations of three nodes and their possible relationships in a DAG [263]. Based on Markov chain rule the probability distributions P for the structures are defined as:

converging connection:	$P(A,B,C) = P(C\|A,B) \cdot P(A) \cdot P(B)$	(5.5)
serial connection:	$P(A,B,C) = P(B\|C) \cdot P(C\|A) \cdot P(A)$	(5.6)
diverging connection:	$P(A,B,C) = P(A\|C) \cdot P(B\|C) \cdot P(C)$	(5.7)

Construction of Bayesian networks

To find the "best describing BN" two steps are applied to construct the networks: *structure learning* and *parameter learning* [263].

Structure learning denotes finding the "highest scoring" structure of a BN. Because this problem is NP-hard for networks, where each node has no more then K parents ($K > 1$) [203] heuristics based on prior knowledge or random selection of interactions are as used to setup an initial structure. Nagarajan et al. propose

three categories to classify the algorithms for structure learning: *constraint-based, scored-based*, and *hybrid* [263].

- *Constraint-based structure learning* is based on pruning the network using the Induction Causation Algorithm as framework for learning/inferring the structures.
- *Scored-based structure learning* follows the idea of assigning a score to the current network and the aim of maximizing the score by altering network structure using Greedy Search, Simulated Annealing, genetic algorithm or similar as long as the network score can be maximized.
- *Hybrid structure learning* defines the combination of both (constraint-based and score-based learning) algorithms.

Parameter learning aims to improve the network structure by altering edges or edge weights after it has been learned. The process is simplified by the Markov property of the networks (each variable is independent of its non-descendants), thus only a limited parental space has to be considered for each node. This is important, because (biological) networks from high-throughput data often have a high number of potential nodes and a small number of observations, which makes parameter estimation complex.
Another approach to reduce complexity is the discretization of the parameters. For parameter discretization, Nagarajan et al. proposed three methods [263]:

- applying of prior knowledge to interpret and assign discrete values directly,
- using heuristics to perform binning of continuous values prior to structure learning,
- testing different intervals and boundaries and choose the best in terms of accuracy and information loss.

To sore the resulting networks and identify the best, the Bayes rule can be applied.

$$P(G/D) = \frac{P(G/D) \cdot P(G)}{P(D)}, \tag{5.8}$$

In equation (5.8) $P(G)$ represents the prior knowledge or any constant non-informative prior. $P(G/D)$ is a function, to be chosen by the algorithm that evaluates the probability that the data D has been generated by the graph G [19]. Popular scores that account for overfitting are the Bayesian Information Criteria (BIC) [194] and the Bayesian Dirichlet Equivalence (BDE) [138].

Dynamic Bayesian networks

One major drawback of BNs is the incapability of modeling loops, although loops are a basic and common structure element of biological networks, (e.g. feedback loops). In 2000, Friedman et al. proposed Dynamic Bayesian Networks (DBNs) to overcome this shortcoming [105]. Figure 5.3 illustrates how circles are modeled by DBNs using time-series expression data. A DAG is generated for each time-point as described for BNs,

while the fusion of these graphs can model circles or loops and results in a graph not required to be a DAG.

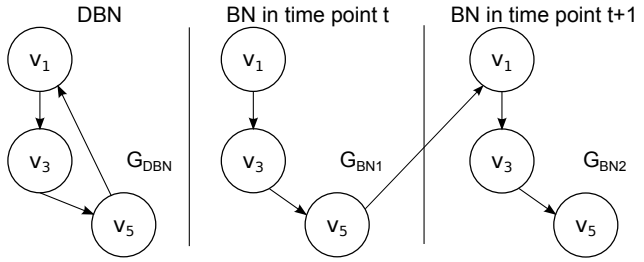


Figure 5.3.: **The left network G_{DBN} represents the Dynamic Bayesian Network (DBN) over two time points of a time-series.** The transitions $v_1 \rightarrow v_3 \rightarrow v_5 \rightarrow v_1$ represent a circle transition over two time points. Right side BN illustrate the DBN on the left $G_{DBN} = G_{BN1} \cup G_{BN2}$, while split into two networks, one for each time-point. The mentioned circle is distributed over both subnetworks (G_{BN1}, G_{BN2}), as the only parent of $v_1 \in G_{BN2}$ is $v_5 \in G_{BN1}$

Alternative approaches proposed to preserve the DAG topology by time-sliced BNs [127, 188]. The basic idea is to prevent circles emerging over time by reducing the network to those variables X_t from a specific time point t and with parents from X_{t-1} only. After merging the networks from different time points to one, identical nodes (genes) $v_1$in different networks are represented as independent nodes $(v_{1_t}, v_{1_{t+1}})$, referenced by the time-point information.

$$v_1 \in \{BN_t, BN_{t+1}\} \rightarrow v_{1_t}, v_{1_{t+1}} \in BN_{ts}, \tag{5.9}$$

where t and $t+1$ represent two time points, ts the time series, and BN_{ts} is the DBN representing both time points. In DBNs, it is required differentiation between, so called intra-slice and inter-slice connections [53]. For example in Figure 5.3, edges v_1, v_3 and v_3, v_5 in graph G_{DBN}, are intra-slice connections, while edge v_5, v_1 is an inter-slice connection, i.e. it connects nodes from different time points $t, t+1$.

5.2.4. Information theory approaches

Information theory networks, relevance networks or mutual information networks are different terms for one class of network reconstruction algorithms. In 1999, Butte and Kohane first mentioned relevance networks to estimate the relation between laboratory tests [45]. In 2000, the same authors adopted this approach to infer relations between genes in expression experiments, as alternative to state-of-the-art correlation analysis [46]. Basic idea is that a high MI of two genes is non-random, rather expressing a biological relation.

To compute the entropy and subsequently the MI, Butte and Kohane suggest discretization of the continuous expression values. As with BNs, the major issue of dis-

cretization is to be sensitive towards signals, since any discretization process implies loss of information [107]. For MI networks, Butte and Kohane suggested a binning approach, where the range of all values is binned using hierarchical clustering and continuous values are mapped to representatives of the specific bins, e.g. bin mean. Meyer et al. [253] propose partitioning all continuous variables V in i bins, where each bin contains/ represents the same number of variables. After discretization, the entropy H for each variable is calculated by:

$$H(V_i) = -\sum_{i=1}^{n} p(v_i) log_2(p(v_i)), \tag{5.10}$$

where $p(v_i)$ is the marginal probability distribution of V_i

Subsequently, the pair-wise MI is calculated and used to infer the GRN. The MI for all pairs of n genes are noted in the so called mutual information matrix (MIM). This square matrix, contains the MI values mim_{ij} for each pair of genes V_i, V_j.

$$mim_{ij} = MI(V_i, V_j) \tag{5.11}$$

The MI is now calculated by

$$MI(V_i, V_j) = H(V_i) - H(V_i|V_j) \tag{5.12}$$
$$= H(V_i) + H(V_j) - H(V_i, V_j), \tag{5.13}$$

where an $MI_{ji} > 0$ means that the two genes are not randomly associated. Where $MI_{ji} = 0$ means that both genes together do not contain more information than one separately.

This results in:

$$MI(V_i, V_j) = \sum_{v_i \in V_i, v_j in V_j} p(v_i, v_j) log_2 \frac{p(v_i, v_j)}{p(v_i)p(v_j)}), \tag{5.14}$$

where $p(v_i, v_j)$ is the probability distribution of V_i and V_j.

As gene expression data is known to follow a Gaussian distribution [259], the equation 5.14 is modified to:

$$MI(V_i, V_j) = \frac{1}{2} log_2 \frac{|C(V_i)| \cdot |C(V_j)|}{|C(V_i, V_j)|}, \tag{5.15}$$

where C is the covariance matrix of the variables and $|\cdot|$ denotes the determinant of C.

5.2.5. Boolean networks

Boolean networks (BoolNs) are one of the simplest approaches to define GRNs. BoolNs are suitable to model interactions and causal relations between nodes. The basic idea is that variable value space is strictly limited to be boolean, i.e. variables in BoolNs can only assume two states, commonly 0 and 1 resp. *false* and *true*. The nodes (genes) in BoolNs are connected by either active or inactive connections (edges). An edge represents logical operator or function that connects two nodes. Thus, a BoolN $G(V, F)$ is defined as a set

of binary variables V (genes) and a set of n boolean functions F (edges between genes). Hence the current state S of the network $V(t)$ and the function $f_i|F$ defines the state of $V(t+1)$, where $t_{1...m}$ are time points or states.

Such networks can illustrate dynamics and kinetics of networks, even without prior knowledge and based on relatively small amounts of input data. Major drawback of the simple and straight forward concept are that BoolNs cannot capture quantities or complex behaviors of real world systems like decreases and increases of gene expression or noise. Required mapping to binary value space makes BoolNs vulnerable for noisy data. Nevertheless, such networks enable time-discrete and synchronous updates and was shown to work well for simulating GRNs [53].

Probabilistic Boolean networks

In 2002, Shmulevich et al. introduced Probabilistic Boolean Networks (pBoolNs) an extension of BoolNs that enables the representation of probabilistic relationships between variables (genes) [325]. The basic idea is to extend BoolNs by assigning a set of function to each node v_i instead of only one.

$$F_i = \{f_j^{(i)}\}_{j=1,...,l(i),} \tag{5.16}$$

where $f_j^{(i)}$ is a possible function that determines the value of gene v_i and $l(i)$ is the number of possible function for x_i.

5.2.6. Ordinary Differential Equations

Ordinary differential equations (ODEs) are used to reconstruct GRNs by describing the expression profiles of each gene by a differential equation over all other genes in the network [19]. The function is defined by:

$$\dot{V}_i = f_i(V_1, ...V_n, Y, \Theta), \tag{5.17}$$

where $V_i = \{v_1, ..., v_m\}$ is the vector of expression of the gene v_i and over all states/times $(s_1, ..., s_m)$, with state-derivates $\dot{V}_i = \{dv_1/ds, ..., dx_M/ds\}$. Y represents external perturbations influencing the system and Θ is the vector of model parameters [274]. The simplest ODE model describes that each expression change of each gene depends on the changes of all other genes:

$$\dot{V}_i = \sum_{j=1}^{N} a_{ij} V_j, \tag{5.18}$$

where a_{ij} denotes the influence of gene v_j towards v_i. If a_{ij} is positive v_j promotes expression of v_i, where negative values of a_{ij} indicate inhibition. Parameter (pre-)setting of a can be utilized to integrate prior knowledge or fixing the model towards further alterations. Popular extension of the models can be achieved by additional parameter that for example simulate external processes like mRNA degradation.

A major problem on applying ODEs for GRN reconstruction is that for the most of the cases, the number of genes is much larger than the number of experiments (small m large p problem). To solve such a differential equation the number of genes must be smaller than the number of experiments [30]. To overcome this problem different approaches were developed, e.g. Gardner et al. [109] assumed that GRNs are unlikely to be fully connected. Therefore, some of the potentially regulating genes p can be set to zero, which reduces the problem. By choosing a $k < n$ the equation can be solved. The k genes that fit the expression data with the smallest error are selected as approximation.

5.2.7. Adaptions of reconstruction algorithms

Over the last years, several extension and combinations of reconstruction algorithms have been developed. For example, classification based approaches like ADANET [328] have been developed to infer relations between genes based on the binary classification. As such, ADANET applies the ensemble classifier AdaBoost [103] to infer GRNs. Huynh-Thu et al. developed Genie3 – an approach that decomposes the problem of reconstruction into regression problems, that are solved by Random Forests or Extra-Trees ensemble methods [160]. Alterations of ODE based methods use approximation (FastNCA [54]) or regression (NIR [109], TSNI [19]) to speed up algorithms and reduce computational cost. Another trend is to combine different reconstruction methods, as for example DELDBN [220]. Thereby, after setting up an initial network structure using DBNs, the GRN is refined by solving the ODE system.

For sake of clarification, it is important to mention that some of the described methods and extension integrate prior knowledge. In practice this has become an effective approach to improve *in silico* network inference methods.

5.2.8. Integration of prior knowledge to GRN reconstruction

Nowadays, the variety of available additional biological information on genomic, epigenomic, transcriptomic, proteomic and metabolomic level is heavily used to assess relevance and influence of regulatory interactions [161, 374]. One of the major aims for using prior knowledge is to reduce complexity, e.g. overcoming the "small m large p problem" by shrinking the number of potential regulators or by assigning parameters like probabilities of potential connections or even precluding specific regulatory interactions [30, 219, 336].

In particular, it is a common strategy to divide the gene set into those, that can regulate others (regulators), like TFs and those, that receive the signal (targets). Defining whether a gene can act either as target or as regulator of an interaction prior network inference helps to reduce complexity and computation effort. For instance, regarding ODEs such information helps to define initial parameter sets. For directed networks like BNs or DBNs such priors reduce the number of allowed interactions [113].

Another common approach is to define an initial network structure. For instance, Zhu et al. applied TFBS and protein-protein interaction (PPI) data to improve network inference using BNs [407].

Prior knowledge can be retrieved from different kinds of sources. High-throughput technologies, genome-wide data sets and the increasing amount of large-scale, multidimensional, and context-specific structured information gives access to biological insides that can be applied to network inference as prior of gene associations [219] or can be used for evaluation of inferred networks. Within this particular study we used additional information and knowledge from databases for the evaluation of the GRN networks (Section 5.5.1). Despite performing experiments by hand, different sources for information and knowledge retrieval have been described and applied for network reconstruction, as for example:

(Online) databases Several databases contain gene regulatory interactions. As there is no common definition for the term regulatory interaction the databases vary in the type of information (e.g. binding, sharing biological context, co-expression, sharing pathways or signaling cascades, experimentally validated physical interaction), size and quality. Many are separated into curated or (experimentally) validated interactions and automatically retrieved or inferred. While for the first group information is often very reliable, the second group lacks on reliability. Important interaction databases are STRING-DB containing experimentally verified and inferred interaction [345], the 5Database of Interacting Proteins (DIP) containing experimentally verified interactions [385] or the Human Protein Reference Database (HPRD) containing manually extracted interactions from the literature [276]. Others contain gene and protein meta information, like the GO [13] or Kyoto Encyclopedia of Genes and Genomes (KEGG) [176], or Pathway data only, like REACTOME [173] or Panther [255]. Databases like Transfac [139] or JASPAR [307] contain information regarding TFs and their target genes, as well as detailed information on the specific binding sites.

DNA sequence analysis Sequence analysis is used to identify TFBS, that indicates potential binding and thus potential regulatory interaction [377] (for details see Section 2.3.2). ChIP-seq experiments are one option to obtain sequences for binding analysis, motif discovery and analysis of co-factorial binding. DNA sequence analysis experiments can be performed or existing experiments can be retrieved from public sequence data repositories, e.g. the European Nucleotide Archive (ENA) [209] or the Sequence Read Archive (SRA) [210].

Literature search In recent years literature search has become popular to search in unstructured data (publications) for co-occurrence of genes. The idea is that two genes potentially interact if they frequently occur in a common context (sentence or paragraph). Some relation extraction methods also aim to estimate from the text, which type of interaction is intended by the text [309, 358, 411].

5.3. Tools implementing GRN reconstruction

Here, we introduce available implementations that could be successfully evaluated using the steady state expression data sets from Chapter 3. 14 out of 26 tools from Table 5.4 could not be tested and evaluated.

Table 5.3 summarizes problems and unmet requirements that led to the individual exclusion. Successfully applied tools are described in the subsequent paragraphs.

Table 5.3.: **Overview of the tools that were discarded from GRN reconstruction comparison.** 11 tools were discarded, because they require time-series data or additional information. Furthermore, no working implementation was found for five tools, while the available implementation of RegnANN is incompatible to required third-party libraries.

Tool	Prior knowledge required	Data requirements	Working implementation
birta	yes	TF-network and miRNA data	yes
BNT	no	time-series data	yes
BoolNet	no	time-series data	yes
deal	required	–	yes
FusGP	no	time-series data	not available
Inferrelator	required	–	yes
MNI	yes	additional expression data	yes
NCA	required	–	yes
REVEAL	no	–	not available
RegnANN	no	–	no, incompatible
SEBINI	no	–	not available
ScanBMA	no	time-series data	not available
TDARACNE	no	time-series data	yes
TSNI	no	time-series data	yes

Weighted correlation network analysis (WGCNA)

Weighted Gene Co-expression Network Analysis (WGCNA) is an implementation of co-expression networks. WGCNA identifies highly correlated (co-expressed) clusters of genes and summarizes these using the module eigengene or an intramodule hubgene. Eigengenes or hubgenes are used for comparison and for calculating membership measures [201]. WGCNA is implemented as R-package and calculates an co-expression matrix for all input genes. The resulting fully connected network can be reduced using thresholds for minimal correlation measures as requirements for establishing an edge. Here, we specified, that two genes are connected by an edge, if their correlation value is greater than the following lower bound. Here, it is implemented as a dynamic threshold for each network:

$$cor_{min} = mean(C_x) + 2 \cdot sd(C_x) \tag{5.19}$$

where C_x is the correlation matrix of input data matrix X. So, an edge is inferred between two genes, if their corresponding correlation values $c_{i,j} > cor_{min}$.

Banjo

Banjo implements static and dynamic Bayesian networks to analyze large data sets. While the Banjo implementation of DNB is intended to infer networks from time-series data, we use the static version as we apply steady state expression data. Furthermore, Banjo offers to integrate known TFs activities and initial network structure for parameter learning. It is important to mention, that Banjo has a number of requirements to the input data. For instance, it is required to discretize the input expression values into intervals $x = 1 \leq i \leq 7$, where $i \in \mathbb{N}$ Banjo comes with two routines (i) binning and (ii) quantile discretization. We applied interval based (binning) discretization using five intervals. As structure learning algorithm we chose simulated annealing. Additionally, the maximum number of parents for a node in the inferred network was set to five. As Banjo extracts a set of potential network with a corresponding score, only the highest scoring network was extracted and evaluated. Furthermore, we removed all self-loop interaction before evaluation.

BNLearn

BNLearn is a R-package that contains six ways of learning Bayesian networks from genes expression data [248]. The six different option/ algorithmic approaches are Grow-Shrink (GS), the Incremental Association Markov Blanket (IAMB), the Fast Incremental Association Markov Blanket (Fast-IAMB), the Interleaved Incremental Association (Inter-IAMB), the Max-Min Parents and Children (MMPC) or the Semi-Interleaved HITON-PC, constraint-based algorithms. Here, we selected IAMB to learn the DAG structure. Thereby the Markov Blanket (MB) for each gene (node) is determined by first identifying all variables that potentially belong to the MB of the current nodes (incl. potential false positives) using mutual information. In the second step, each variable in the MB is tested whether it is independent regarding the other variable of the MB [360]. After performing the structure learning loops were removed.

BC3NET

The R-package BC3NET [83] is an extension of the mutual information (MI) based C3NET algorithm by Altay and Streib [9]. Basic idea of both is to identify the "conservative causal core", i.e. the statistical significance MI network of all pairs of genes. In detail, two genes are connected if their shared MI is maximal for at least one of both with respect to all other genes. The C3NET algorithm calculates the MI for all pairs of genes before the edges that do not satisfy the mentioned constraint are iteratively removed . BC3NET extends the general principle of C3NET by bootstrap aggregation called "bagging" to create an ensemble of data subsets before for applying C3NET to each of the subsets [82]. Finally, the resulting networks from all subsets are aggregated using statistical hypotheses testing to determine a threshold parameter for the aggregation.

C3mtc

C3mtc is another extension of the C3NET algorithm (see BC3NET) that aims to implement a more efficient strategy of edge selection and adds a final step of multiple testing correction to account for the type one error [83]. While step one is similar to C3NET (calculating MI for all gene pairs), the second step is an extremal selection strategy, allowing only one outgoing edge per gene, to reduce the amount of interaction to be tested. Finally, in the third step multiple testing correction is performed to obtain the relevance for the interactions.

NIR

Network Identification by multiple Regression (NIR) belongs to the category of ODE-based methods and estimates regulatory interactions based on multiple linear regression [109]. The aim is to create a model from RNA expression changes by using a set of steady state transcriptional perturbations. The method was originally developed to estimate the gene regulatory effects during treatments. It follows the assumption, that changes in steady state data are mainly influenced by the treatment, rather then by cell-development or other time-depended effects. Here, we circumvent this assumption by considering the different cell type (iTreg/ nTreg cells) as treatment differences and trying to estimate regulatory interactions between both. Additionally, we grouped the microarrays regarding their original study correspondence to calculate the individual standard error for each study, which is used to estimate the significance of gene expression changes.

ARACNE

Since in 2006, Margolin et al. presented the "Algorithm for the Reconstruction of Accurate Cellular Networks" ARACNE [249], it has become one of the most popular and most often adapted GRN inference tools. ARACNE uses MI to determine the most statistical significant relation between two genes that can not be explained by errors, artifact of other statistical dependencies in the network. It assumes, that such irreducible statistical dependencies is likely to identify direct regulatory interactions. The algorithm calculates the MI between all pairs of genes as well as the information-theoretic measure of relatedness. A major drawback of that approach is, that it is not sensitive to mediated relations, i.e. genes separated by one or more intermediaries may be highly co-regulated without implying an irreducible interaction. Thus ARACNE infers a direct interaction between two genes, if there are connected via intermediates. To overcome this, the Data Processing Inequality (DPI) is used to remove such false positive indirect interactions. The DPI states that in a clique of three genes (g_1, g_2, g_3) where two genes (g_1, g_3) interact only through the third gene g_2, then $MI(g_1, g_3) \leq min[MI(g_1, g_2); MI(g_2, g_3)]$. Thus the least of the three MIs can come from indirect interactions only and is removed.

Here, we applied two extension of ARACNE implemented by the R-package "parmigene" [305]. ARACNE.a implements an adaptive model, while ARACNE.m implements a multiplicative model to estimate the significance of the MI.

MRNET

MRNET is a "Maximum Relevance Minimum Redundancy" (mRMR) based network inference approach belongs to the class of information theory approaches [275]. It calculates the MI for all gene pairs and uses the feature selection approach mRMR to identify the "relevant" interactions in the network, while discarding the rest.

CLR

Context Likelihood of Relatedness (CLR) is another machine learning approach that uses the concept of information theory networks [97]. After calculating the MI information for each pair of genes, CLR applies a correction step to eliminate false correlation and indirect influences (see ARACNE) by calculating the statistical likelyhood of the pair-specific MI in the context of the "background" distribution. Background distribution of MI denotes the distribution of MI over all pairwise interaction partners. Only those MI that are significantly higher, than the "background" are considered.

Genie3

A regression tree-based approach of decomposing the problem of network inference is implemented by Genie3 [160]. Thereby, the problem of network inference for n genes is decomposed into n regression problems, where the aim is to predict the expression profile of one gene by all others. The regulatory importance of a gene towards a target gene is collected over all sub-problems and ranked to create a general regulatory network for the data sets. More general the approach of Genie3 is to perform tree-based feature selection to select those genes that have highest descriptive power towards the expression pattern of the other genes. Genie3 is implemented as a R-Script.

5.4. Overview of selected GRN reconstruction methods and tools

Table 5.4 gives an overview about the advantages and disadvantages of the described methods and lists for each method a set of available implementation. To the best of our knowledge, there are no implementations using artificial neural networks (ANNs) currently available. Therefore this class of network reconstruction algorithms was not described and considered in this work.

Various authors compared and summarized algorithmic approaches for GRN inference before, e.g. [53, 125, 234, 296]. As in this work too, many reviews and comparisons are focused on GRN inference from either steady state or time-series data only. Chai et al. [53] or Hempel et al. [141] focused on methods for inference based on time-series data. Others analyzed and compared approaches without focusing on temporal aspects, like in [19, 108, 137, 247, 296]. Hecker et al. [137] formulated two important questions for benchmarking reconstruction algorithms: (i) Does the model correctly predict the behaviors of the GRN? (ii) Does the model represent the true structure of the system?.

Table 5.4.: **Methods and Tools for GRN inference.** For each method/class of methods (upright written in most left column) we list a number of tools. Underlined tools are used for network inference in Section 5.5.3. Tools marked with an asterisks are not longer supported or not available.

	Advantages	Disadvantages	Tools
Coexpression/ clustering networks	- Computational low complexity - Scale for large networks - Infer cyclic networks	- Infer undirected connections - No noise handling - Handle continuous values - Infer complete networks or require threshold setting	WGCNA [201] Hierarchical Clustering [92]
Bayesian networks	- Handle noise and uncertainties - Able to work on the logically interacting components with small number of variables - Integrate the prior knowledge to strengthen the causal relationship - Infer the structure of network statistically	- Infer undirected connections - Do not scale for large network (large number of nodes) - Cannot consider temporal information from time-series data - Do not support (feedback-) loops	Banjo [398] BNLearn [316] SEBINI* [353] birta [401] deal [38]
Dynamic Bayesian networks	- Model (feedback-) loops - Handle perturbation and structural modification of the network - Model direct and indirect causal relations - Consider temporal flow of time-series data	- Computationally slow - Do not scale for large network (large number of nodes) - Sensitive to missing values	BNT [260] ScanBMA [396] FusGP [258] Banjo [398] REVEAL* [221]
Boolean networks	- Scale for large networks - Only binary variables simplify simulation - Easy to interpret	- Computationally slow - Sensitive to missing values - Require data discretization	BoolNet [261]
ODE networks	- Handle steady state and time-series data - Easy to change/ adapt - Infer directed, signed networks - Return quantative measure of effect	- Require $\|genes\| \leq \|experiments\|$) - Computationally intense - Parameter search requires priors	NCA [222] TSNI [19] NIR [109] MNI [85] Inferelator2 [243]
Neural networks	- Handle steady state & time-series data - Can infer "any" functional relationship incl. loops, cyclic graphs, etc. - Infer directed and undirected networks (specific method dependent) - Robust regarding noise	- Computationally complex (slow) - To avoid over-fitting training is required on different data sets - Only for small networks due to computation cost	RegnANN* [120]
Information Theory based networks	- Easy integration of prior knowledge - Capture non-linear interactions - Fast computation	- Threshold estimation necessary to avoid complete networks - Require data discretization	ARACNE [249] TDARACNE [410] C3NET [9] Genie3 [160] MRNET [254] CLR [97]

The first question refers to the (model) output of the inference and its accordance with the actual observed outcome from wet-lab experiments. The second question depicts the specific gene interactions of the model, i.e. the truth or correctness of each inferred interactions is of importance. To answer the second question detailed knowledge of the

organism and the true regulatory interactions, as it is the case for model organisms like *E. coli* or *S. cerevisiae* is required.

As mentioned, the most popular benchmarking setup for GRN reconstruction based on gene expression data was developed in 2010 as part of the DREAM5 network inference challenge [247]. Thereby, the reconstructed GRNs were evaluated using binary classification using standard performance metrics from machine learning (precision vs. recall (PR) and receiver operating characteristic (ROC) curves). Marbach et al. also evaluated the power of combining different algorithms to asses more robust networks, by stacking the approaches [247]. Thereby, they observed, that no single method performs very well over all datasets, while the combination of tools achieved better and more robust results.

5.5. Application of reconstruction tools to infer a Treg cell specific GRN

In this section we applied a subset of the described tools to the prior described Treg cell meta expression set (Chapter 3). Thereby, we aim to compare the performance of different GRN reconstruction approaches. Tools from only four out of the seven categories (listed in Table 5.4) were tested. Before we describe the setup of the evaluation strategy, we introduce the public data repositories that we used as evaluation databases. After applying the selected reconstruction tools, we report and compare their performances.

5.5.1. Evaluation databases (silver standards)

To evaluate whether the inferred edges are known gene-gene interactions or not, we compare them to the set of approved interaction from two major interaction databases, namely ConsensusPathDB (CPDB) [175] and STRING-DB [102]. Both databases do not claim to contain the complete set of existing interactions. In addition, both focus on integrating different source and assign confidence levels (e.g. predicted, mined from literature, experimentally validated) to the interactions. It is important to mention, that we use domain independent databases for evaluation, because no T cell or immune cell specific interaction databases or repositories were not freely and publicly available or as in the case of InnateDB are part of CPDB or STRING-DB.

ConsensusPathDB (`http://consensuspathdb.org`) is an integrated database that combines information like protein interactions, signaling reactions, metabolic reactions, gene regulations, genetic interactions, drug-target interactions and biochemical pathways from 30 public resources databases into a single schema and thus grants access to more than 150,000 unique physical entities and 500,00 interactions, including about 17,000 gene regulatory interactions. For the presented evaluation strategy we downloaded and prepared the murine protein-protein interaction set and pathway set from CPDB that contains manually added interactions as well as interactions from the following sources:

BIND The Biomolecular Interaction Network Database contains protein function information, interaction and pathway data [15],

Biogrid Biological General Repository for Interaction Datasets is a interaction repository with data compiled from comprehensive manual curation [334],

PDB The Protein Data Bank mainly focuses on information about the 3D folding structures of proteins [28],

IntAct IntAct Molecular Interaction Database provides interaction from literature curation and manual submission [182],

MINT The Molecular INTeraction database contains structured information about gene and protein interactions [267],

DIP Database of Interacting Proteins is a database of experimentally validated protein interactions [306],

InnateDB is a systems biology data repository for interactions and pathways related to the innate immune response [40],

MatrixDB a curated database of experimentally established protein data [57],

MIPS-MPPI the MIPS Mammalian Protein-Protein Database is a collection of manually curated high-quality PPI data collected from the scientific literature by expert curators [271].

ConsensusPathDB data integration and mapping was performed using UniProt and Entrez gene identifiers.

For the evaluation we used two subsets of the full CPDB, (i) all protein-protein interactions (CPDB.ppi) and (ii) pathway information for mouse (CPDB.pw). For (i) we used the provided identifier mapping files to translate interactions from UniProt ids to ensembl mouse gene ids, while for the second CPDB provides a list containing ensembl mouse gene ids. (i) contains 17,089 PPIs of mouse proteins and (ii) lists 2,141 pathways.

STRING-DB database (`http://string-db.org`) is one of the largest public interaction databases, today. It contains roughly one billion interactions for more than 2,000 organisms annotated with confidence levels. A confidence score assigned to each interaction is used to indicate and verify the level of validation. STRING distinguishes five main sources of interactions: (i) genomic context predictions, (ii) high-throughput lab experiments (iii) (conserved) co-expression, (iv) automated textmining, (v) previous knowledge in databases [102]. In 2012, STRING-DB contained about two million PPIs for mouse. A confidence score is assigned to each interaction, that corresponds to the probability of finding the linked proteins within the same KEGG pathway [369]. As many of the included PPIs have a very low confidence score, we use two subsets for evaluation, (i), only those interactions with are confidence score of at least 400 (STRING.400) and (ii) the set of experimentally validated interactions, only (STRING.exp).

Using information from CPDB and STRING-DB we obtained four evaluation sets. Additionally, we created a set by integrating all sets into one, which is indicated as "combined".

Regarding the subsequent application, an edge is considered as existing (true), when it is either listed as an interaction in STRING.exp, STRING.400, CPDB.ppi, combined,

or both genes share a common pathway from CPDB.pw or combined.

5.5.2. Tool testing strategy

The selected GRN reconstruction tools are applied to the test strategy (Figure 5.4), which consists of six steps, described as follows: As input data for testing the strategy we used the steady state Treg cell gene expression data set, which was presented in Chapter 3.

1. **Preprocessing**: The microarray experiments from the selected expression data set are formatted and transformed regarding to the tool specific requirements (see tool descriptions in the following Subsection 5.3). Because some tools do not allow continuous expression values, these must be discretized or binned. Others require special input formats like matrix representation.
2. **Inference**: This step depicts the actual inference of the applied tools. Inference is performed on ten gene sets. Each tool is tested ten times while each gene set constitutes an initial expression set. The sets are incrementally built on each other, i.e. the set of 20 genes contains the set of 10 genes, and so on and so forth.
3. **Edge extraction**: The tools specific output format are translated into an adjacency list format for subsequent performance estimation.
4. **Evaluation**: Evaluation depicts the validation whether the inferred edges are present in the selected evaluation sets or not. During evaluation we apply the two gene-interactions databases and four filter sets. For each tools we extract the number of true positively, true negatively, false positively, and false negatively inferred interactions. The direction of interactions was not considered during evaluation.
5. **Performance estimation**: Performance measures (accuracy, specificity, precision, recall and F-measure) are calculated for each method based on the results from evaluation step.
6. **Ranking**: In this step tools are ranked regarding their performance measures.

5.5.3. Results – Performance of GRN reconstruction tools on Treg cell expression data

Here, we describe the results of the network inference by the reconstruction tools towards the earlier presented Treg-subtype specific data set generated from meta-analysis. After data preprocessing, inferrence and result formatting we obtained an adjacency list of inferred interactions for each tools and each gene set.

In Table 5.5 we present the number of inferred interactions per tools based on the supplied gene set and additionally the number of known interactions in evaluation data sets based on the gene sets. The numbers illustrate the differences in the size of the inferred networks, which in the end also results from the individual thresholds applied for the tools. For example the exclusion of loops/ties tremendously reduced the number of inferred interactions for Banjo and WCGNA. Furthermore, lower bounds for edge weights, like significance levels had to be applied to the tools WGCNA, BC3NET, C3mtc,

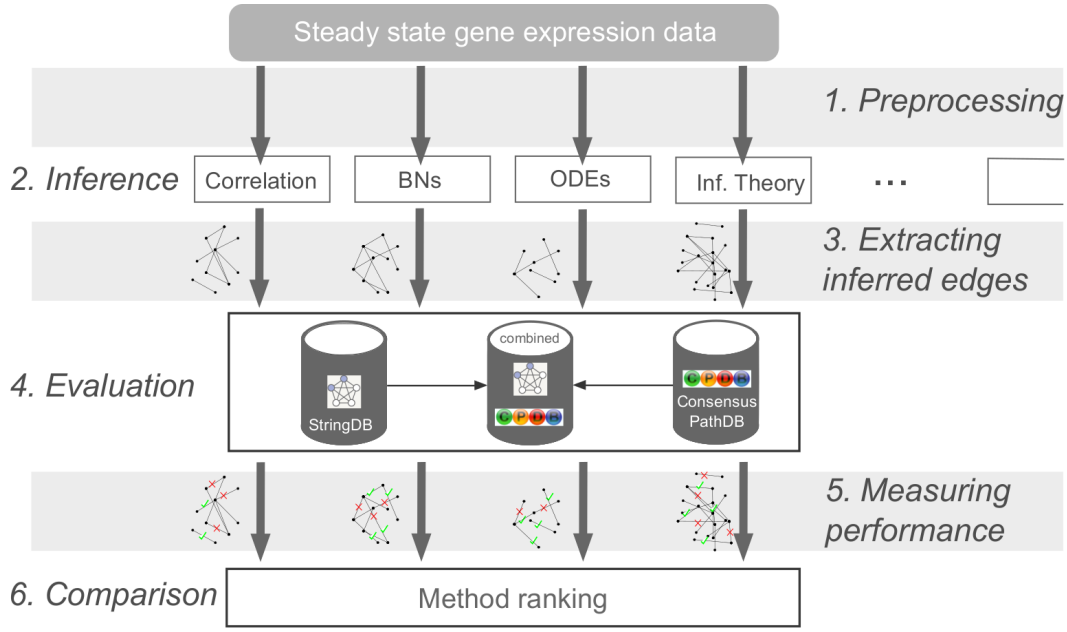


Figure 5.4.: **Tool testing strategy.** The six layers (1-6) illustrate the main steps of the pipeline in processing order.

ARACNCE.m, ARACNE.a to prevent inferring complete networks, while a large fraction of the interactions have very low predicted regulatory influence. Since the number of inferred interactions depends on individual thresholds we do not interpret the size of the inferred networks, rather than testing how many of the inferred interactions can be found in one of the five evaluation data sets.

After applying the presented test strategy we assessed individual tool performances using the following metrics:

True positive (TP) depicts the number of correctly inferred edges.

False positive (FP) depicts the number of inferred edges that do not exists in the true network.

True negative (TN) depicts the number of correctly not inferred edges.

False negative (FN) depicts the number of edges that exists in the true network but were not inferred.

Accuracy is the proportion of correctly inferred interactions (true positives) plus the correctly not inferred interactions (true negatives) to all interactions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Specificity/True negative rate is the proportion of correctly not inferred interactions

Table 5.5.: **Number of interactions inferred by the presented tools (rows) tested on ten gene sets (columns).** The table header (bold) shows ten genes sets, the column names indicate the number of genes each set contains. The table is horizontally divided into two parts the upper part lists for each applied tool (rows) the total number of inferred edges per gene set. The lower horizontal part "Evaluation data sets" shows for each evaluation set the number of existing edges based on the given gene set.

	Gene set size	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
Tools	ARACNE.a	17	36	62	94	136	187	230	268	329	374
	ARACNE.m	19	59	103	147	212	274	366	443	541	619
	Banjo	7	16	25	36	48	58	87	89	106	127
	BC3NET	11	26	48	63	80	100	124	158	183	224
	BNLearn	14	32	38	53	55	69	61	60	75	71
	C3mtc	9	18	25	32	40	49	59	67	75	86
	CLR	34	130	280	489	766	1107	1525	1980	2507	3099
	Genie3	17	27	45	60	62	76	116	133	76	78
	MRNET	28	110	263	473	745	1079	1499	1940	2452	3037
	NIR	30	60	90	120	150	180	210	240	270	300
	WGCNA	2	7	15	26	36	55	72	91	125	154
Evaluation data sets	STRING.exp	4	22	58	102	138	162	172	212	272	340
	STRING.400	10	52	112	184	246	316	380	458	578	716
	CPDB.ppi	5	10	19	23	30	34	41	45	48	50
	CPDB.pw	29	124	260	476	610	869	1017	1290	1915	2622
	combined	34	137	293	520	670	942	1120	1413	2059	2777

among all truly not existing interactions.

$$Specificity = \frac{TN}{TN + FP}$$

Precision is the proportion of correctly inferred interactions out of the full set of inferred interactions.

$$Precision = \frac{TP}{TP + FP}$$

Sensitivity/Recall/True positive rate is the proportion of correctly inferred interactions to all interactions.

$$Sensitivity = \frac{TP}{TP + FN}$$

F-measure / F1 is the harmonic mean of precision and recall.

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Table 5.6 shows the results for all applied tools (rows) on the different evaluation data

sets (columns).

Based on meta expression data set from Chapter 3 we created ten increasing subsets containing 10 to 100 genes. The gene sets composed as follows: first, we filtered for all genes for TFs and Cytokines which have been defined as relevant for T cell regulation by Wei et al. [378]. Second, the resulting genes were tested for significant differential expression and ranked regarding their fold changes (between iTreg and nTreg cell experiments). The ten genes with the highest fold changes were assigned to the first evaluation set, the 20 genes with the highest fold change to the second and so on and so forth. In consequence, the smaller sets are part of the large ones.

To provide an overview about the average performance of the tools we calculated the mean over all data set sizes (10 to 100 top genes) Table 5.6 illustrates, that almost all tools do not infer "known" interactions indicated by the low sensitivity rates (TPR) of lower 1pp. Nevertheless, the majority of inferred interactions is correct for nine of 11 tools, as accuracy values above 0.5 indicate for all evaluation data sets. Those metrics in Table 5.6, that are marked with an asterics must be considered very carefully, as the corresponding tools (Banjo and Genie3) were not able to infer any correct interaction except for self-loops (that were removed prior evaluation). Their sensitivity value of 0 corrupts all other metrics for these tools. But, accuracy values mainly result from high specificity, which is high because of the low number of inferred interactions for all tools (see Table 5.5). Because not inferred interactions count as negative and the inferred networks are rather small specificity (true negative rate) is high. This is illustrated in Table 5.5 by the number of inferred interactions per tool and the number of known interactions for each evaluation data sets. The number of theoretically allowed interactions ($n^2 - n$ for directed and $n * (n - 1)/2$ for undirected networks) is very large in contrast to the number of proven interactions included in the databases. Furthermore, also the number of inferred interaction heavily varies between the tools.

Comparing sensitivities illustrates that the tools are not able to infer large fractions of existing interactions. While specificity is high due to the high number of true negative (TN) inferred edges, as this is high by chance as the most possible edges are negative in the evaluation networks, too. To account for this effect we also calculate specificity values for all tools and gene sets (illustrated in Figure 5.5) as well as an average F-measure and accuracy illustration over all gene sets in Figure 5.6. Both illustrate the drawback of inferring large numbers of interactions, as specificity drops if many false positive (not existing interactions) are inferred.

In detail, CLR und MRNET, two machine learning based information theory approaches show highest F-measures, while their specificity and accuracy is the worst among the tested tools (see Figure 5.5). This relies on the fact, that both methodw infer by far the most edges, which in turn reduces the amount of correctly inferred negative interactions (specificity). NIR performed best in terms of accuracy, which is mainly founded by the high specificity.

Finally, results show, that none of the presented tools is able to infer a relevant set of known interactions. The results are insufficient for further analysis as the fraction of false positives among the very low number of correctly inferred interactions is high. So far we can not estimate, whether a significant number of the inferred interactions is true

Table 5.6.: **Performance of GRN reconstruction tools tested using five evaluation set.** The table is vertically divided into four parts, each for one metric, namely sensitivity, specificity, accuracy and F-measure. Tools are shown in the rows of each part, while the columns present the evaluation data set. Best performing tool per metric is highlighted in bold. Numbers marked with an asterics(*) must be interpreted carefully, see text.

	STRING.exp	STRING.400	CPDB.ppi	CPDB.pw	combined
			Sensitivity (TPR)		
ARACNE.a	0,020	0,067	0	0,128	0,146
ARACNE.m	0,024	0,079	0	0,196	0,215
Banjo	0	0	0	0	0
BC3NET	0,010	0,023	0	0,094	0,098
BNLearn	0,009	0,020	0	0,073	0,076
C3mtc	0,004	0,014	0	0,059	0,060
CLR	**0,072**	**0,201**	**0,001**	**0,517**	**0,586**
Genie3	0	0	0	0	0
MRNET	0,071	0,190	**0,001**	0,495	0,554
NIR	0,018	0,024	**0,001**	0,069	0,072
WGCNA	0,006	0,012	0	0,045	0,047
			Specificity (TNR)		
ARACNE.a	0,805	0,827	0,796	0,855	0,863
ARACNE.m	0,712	0,738	0,701	0,791	0,800
Banjo	*1**	*1**	*1**	*1**	*1**
BC3NET	0,873	0,878	0,868	0,914	0,916
BNLearn	0,883	0,888	0,879	0,916	0,917
C3mtc	0,921	0,925	0,919	0,949	0,949
CLR	0,104	0,162	0,075	0,295	0,326
Genie3	*1**	*1**	*1**	*1**	*1**
MRNET	0,162	0,214	0,133	0,342	0,367
NIR	0,929	0,930	0,926	0,939	0,940
WGCNA	**0,955**	**0,957**	**0,953**	**0,972**	**0,972**
			Accuracy		
ARACNE.a	0,580	0,609	0,567	0,647	0,659
ARACNE.m	0,517	0,552	0,502	0,624	0,636
Banjo	*0,853**	*0,853**	*0,853**	*0,853**	*0,853**
BC3NET	0,623	0,631	0,617	0,678	0,680
BNLearn	0,632	0,639	0,626	0,675	0,677
C3mtc	0,655	0,661	0,652	0,691	0,692
CLR	0,100	0,178	0,059	0,364	0,406
Genie3	*0,853**	*0,853**	*0,853**	*0,853**	*0,853**
MRNET	0,138	0,209	0,097	0,387	0,422
NIR	**0,797**	**0,799**	**0,791**	**0,814**	**0,815**
WGCNA	0,677	0,680	0,673	0,700	0,700
			F-measure (F1)		
ARACNE.a	0,029	0,089	0	0,172	0,196
ARACNE.m	0,029	0,093	0	0,230	0,253
Banjo	0	0	0	0	0
BC3NET	0,016	0,036	0	0,140	0,147
BNLearn	0,015	0,032	0	0,106	0,110
C3mtc	0,008	0,024	0	0,095	0,097
CLR	**0,045**	**0,126**	0	**0,321**	**0,365**
Genie3	0	0	0	0	0
MRNET	**0,045**	0,124	0,001	0,320	0,358
NIR	0,025	0,034	0,001	0,093	0,097
WGCNA	0,012	0,022	0	0,082	0,084

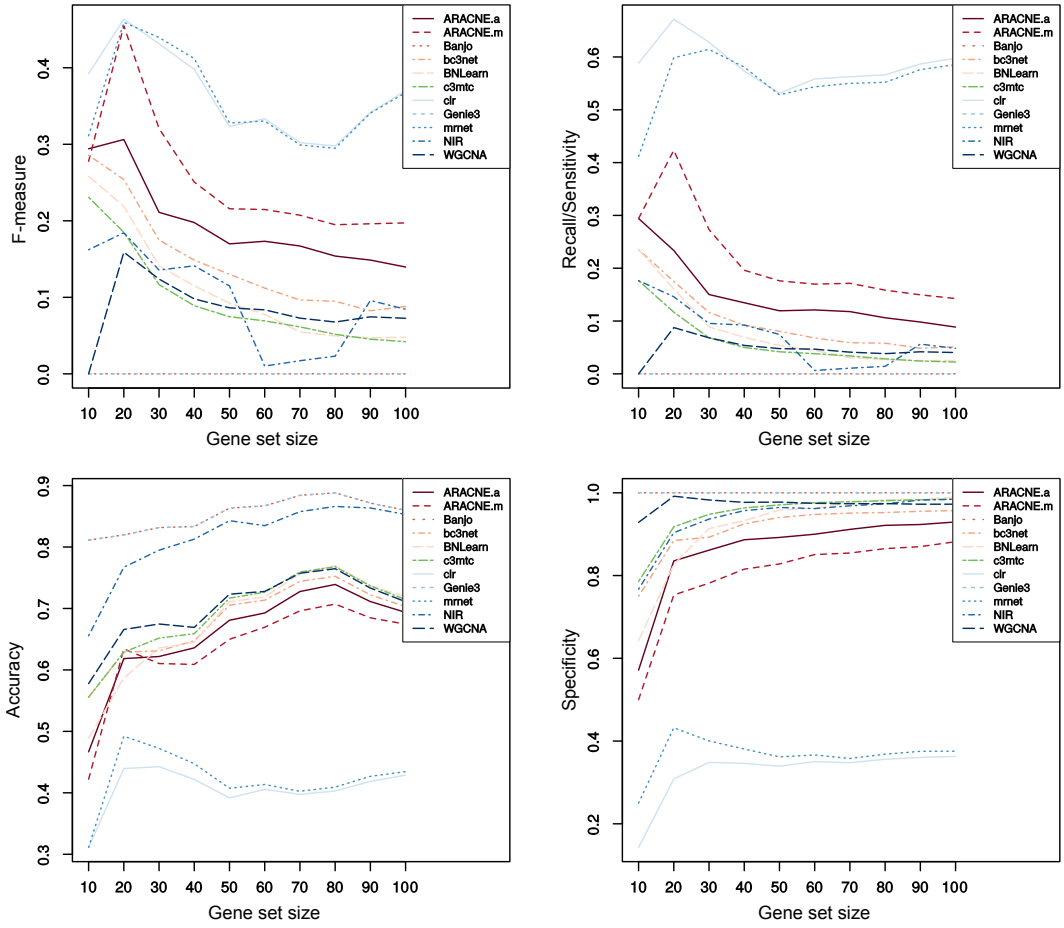


Figure 5.5.: **The line plots illustrate the tool specific characteristics of the performance metrics (y-axis) for increasing gene set sizes (x-axis).** From upper left to lower right the figures show F-measure, sensitivity, accuracy and specificity. Labels on the x-axis depict the gene set size. The measures were calculated based on the "combined" evaluation corpus. The lines for Banjo are plotted before the lines of Genie3 and thus fully or almost invisible.

but missing in the evaluation databases, or not.

Nevertheless, comparing the top three most often inferred interactions over all gene sets revealed three edges: *Ccla4-Lta*, *Lta-Il2*, *Il2-Foxp3*. *Ccla4* is a chemokine which is known to be functionally different in different T lymphocyte subsets and specific for CD8 lymphocytes [227]. The protein family associated to the cytokine *Lta* regulates Treg suppressor function *in vivo* but not *in vitro* [41], which is interesting as our data compares iTreg with nTreg cells. *Il2* and *Foxp3* are two well known key regulators for T cell. While *Il2* is the most important cytokine during Th2 cell induction it is also important for the signaling during the development and function of regulatory T cells [244]. *Foxp3* is the

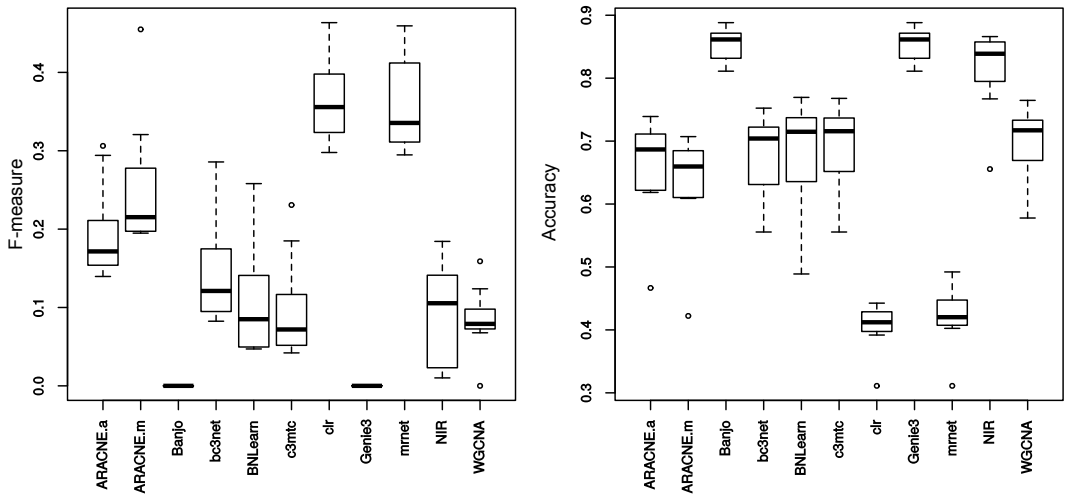


Figure 5.6.: **Boxplots of F-measure and accuracy aggregated for each tool over all gene sets to illustrate variance.** Comparison shows that no tool is able to outperform all others in terms of F-measure and accuracy. Even more we observed that higher number of correctly inferred interactions come along with a higher number of false positives.

master transcription factor for Treg cells (see Chapter 2). Furthermore it is known that a reduction in Il2 transcription results in impaired FoxP3$^+$ CD4$^+$ Treg-cell recruitment and function [318].

5.5.4. Discussion of the results from GRN reconstruction

The presented application of GRN reconstruction methods using the expression data set derived from meta-analysis did not reconstruct large fractions of known interactions and thus encouraging results for further research. Table 5.5 shows that the number of known interactions (interactions found in evaluation sets) is rather small, compared to the theoretically possible number of interactions. Even worse, it is impossible to estimate the number of truly not existing interactions, since a gold standard is missing. We illustrate this inadequacy for the chosen evaluation sets in Table 5.7. The table shows the number of theoretically possible edges and the number of corresponding edges for each evaluation sets.

The comparison of the inferred interactions between the tools also revealed, that only a small fraction of interactions is commonly inferred by multiple tools. Table 5.8 indicate the low overlap of interactions. Marbach et al. state in [247] that a high overlap between different tools increases the reliability of the inferred interactions. In this study we could not identify large overlaps of inferred networks. Table 5.8 shows that the overlap of identically inferred edges by six or more tools is very low. Moreover, tools based on a similar or common theoretical approach tend to infer more identical edges, e.g. MRNET, CLR, and both ARACNE implementations. The low amount of overlapping

Table 5.7.: **Estimation of potential edges per network.** Here, we compare the maximal possible number of edges per network ("full, directed, without loops") with the number of corresponding edges in evaluation sets.

Gene set size	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
full, directed net	90	380	870	1560	2450	3540	4830	6320	8010	9900
STRING.exp	4	22	58	102	138	162	172	212	272	340
STRING.400	10	52	112	184	246	316	380	458	578	716
CPDB.ppi	5	10	19	23	30	34	41	45	48	50
CPDB.pw	29	124	260	476	610	869	1017	1290	1915	2622
combined	34	137	293	520	670	942	1120	1413	2059	2777

inferred edges give reason to disregard further ensembled analysis using majority voting, i.e. constructing a network based on those edges inferred by the majority of tools. Furthermore, applying and testing the union of the inferred networks as additional approach was neglected, because of the high false positive rates (FPRs) and the low overlap of inferred interactions.

Table 5.8.: **Overlap of inferred edges per gene set.** The rows represent the genes sets (10 to 100), while the columns represent the frequency of inference. The column "in total" contains the number of inferred unique edges. For example the column 2× depicts the number of overlapping edges that were inferred by exactly two tools, where 11 indicates how many unique edges were inferred by all applied tools.

		Unique interactions inferred										
		in total	**1×**	**2×**	**3×**	**4×**	**5×**	**6×**	**7×**	**8×**	**9×**	**10×**	**11×**
Gene set size	10	*71*	29	13	7	9	5	5	3	0	0	0	0
	20	*208*	68	56	37	23	15	5	1	2	0	1	0
	30	*417*	131	139	63	45	29	5	2	1	1	1	0
	40	*679*	185	278	90	84	22	12	4	1	2	1	0
	50	*994*	228	470	126	109	39	11	5	3	2	1	0
	60	*1388*	289	720	151	154	43	11	11	4	4	1	0
	70	*1880*	352	1031	211	202	43	21	12	3	5	0	0
	80	*2399*	422	1389	253	249	40	22	15	4	5	0	0
	90	*2945*	470	1746	321	307	54	24	15	5	3	0	0
	100	*3594*	532	2230	371	338	67	28	20	3	5	0	0

Regarding potential improvements in the presented pipeline, we focus on two steps:

1. **Identifying proper gene sets for inference**
   Gene set selection can be improved by filtering for genes based on common expression patterns or gene enrichment properties. Rather than filtering genes by the fold change ranking, clustering based on expression data to identify common

expression patterns is an alternative. Additionally, enrichment analysis tools like GSEA [343] or GO term analysis [98] can be used to create more specific gene sets e.g. based on shared meta information.

2. **Extending GRN reconstruction by prior knowledge integration**
   Another, very simple way to achieve better results is to make use of the tools functionality to include prior knowledge, as it has been show in a number of studies like [12, 29, 113, 119, 161, 281, 335, 381]. For instance, sources of prior knowledge can be databases like those we used for evaluation or ideally specific additional experiments like ChIP-seq to determine potential TF activities and binding partners [20]. While some of the tested tools (e.g. Banjo, NIR, ARACNE) optional allow, other tools require the integration of additional prior knowledge (and therefore have not been considered in this study - see Table 5.3).

We choose to perform the analysis without additional knowledge about known regulators or functional information, as not all tested tools provide this functionality. Some of the presented tools allow the integration of prior knowledge, as for example, Banjo or ARACNE can use initial network structures, while others integrate gene information like "is a transcription factor" to define sets of regulators. To keep evaluation and results interpretation less complex, we refrained from providing different prior information to the tools.

We aimed to assess tool performances on "real", noisy data from aggregated, heterogeneous, publicly available experiments. It was clearly shown, that the tools do not provide high inference quality based on such data for known gene-gene interactions. The presented study pointed out, that at least a comprehensive gene selection strategy and prior knowledge is required to increase network inference quality.

For the field of T cell lymphocytes, studies of Ciofani et al. [66] and Äijo et al. [4] showed the high impact of GRN reconstruction and additional knowledge integration to investigate regulatory events during T cell activation and T cell fate decision. In contrast to our work, in both studies time-series data was applied to assess the regulatory networks. Interestingly, Ciofani et al. reconstructed a Th17 cell specific GRN using time-series RNA-seq and ChIP-seq data and applied the Inferelator tool (see Table 5.4) to infer a Th17 cell specific GRN. To show the impact of external information we present a GRN reconstruction approach for Th2 cell exclusively based on the data integration from different experiments and database information in Chapter 6.

5.6. Summary

In this chapter, we presented the application of GRN reconstruction algorithms and tools to a large set of heterogeneous Treg cell specific gene expression data. A collection of tools and algorithmic approaches from different categories were presented and compared regarding their advantages and their disadvantages (Table 5.4). We presented a general GRN inference strategy, compared and tested a selection of reconstruction tools on the Treg subtype specific meta-expression data set from Chapter 3. The presented strategy

did not incorporate prior or additional knowledge, therefore the range of available tools was limited. Other tools were excluded because, no functioning implementation was available. Each of the finally selected GRN reconstruction tools was applied to ten gene sets of varying size (number of genes). The gene sets contained those with the highest fold changes between both Treg subtype in the differential expression analysis of the meta-expression experiment. The presented test strategy also includes an evaluation phase, within the inferred networks were compared to information from public interaction databases, namely STRING-DB and ConsensusPathDB.

In the end, we inferred 110 networks which were compared regarding known interactions. The results do not show high consistency of inferred edges between the tools, as only small fractions of edges were inferred by multiple tools, furthermore, tools with overlapping sets are very similar regarding their baseline algorithmic approaches.

Finally we can summarize, that the reconstruction of a GRN from assembled publicly available gene expression data, that obtains high heterogeneity, does not reveal networks with high reproducibility using different inference methods. Nevertheless, under the large amount of unknown interactions, we could infer Treg cell related and experimentally validated interactions using the majority of tools since the analysis of the top-most often inferred interactions over all tools shows Treg cell specificity.

6. Data integration for constructing T cell subtype specific regulatory networks

In this chapter we describe the advantages of integrating multiple *OMICS* data sets to construct a specific, small-scale, yet high resolution GRN. *OMICS* refers to data from different kind of biological experiments, as for example, transcriptomics, genomics, proteomics. Here, we used publicly available Th2 cell specific data sets. The generated data was obtained from experiments performed by the group of Ria Baumgrass from the German Rheumatism Research Center Berlin. The integration strategy in this work combined gene expression data from microarray, RNA-seq, and DNA-seq data into a single GRN, centered around *Stat6* – the Th2 cell master TF. The work in this chapter describes a collaboration with the German Rheumatism Research Center (DRFZ) and the group of Joachim Selbig at the University of Potsdam. It was published in 2016 [168]. The authors contribution to this work was the development of the overall integration strategy, data source specific analysis workflows for the ChIP-seq analysis, RNA-seq data analysis of time series data, and microarray data analysis. Furthermore, the author contributed to the network visualization using Cytoscape [319].

6.1. Motivation

Here, we describe how data integration can help to reconstruct high-resolution GRNs. The availability of *OMICS* data from high-throughput experiments keeps increasing every year (see Figure 3.1, p.24) and provides comprehensive descriptions of nearly all molecular components and interactions within the cell [174]. We aim to integrate such information from matching and complementary experiments like time-series RNA expression, TF-binding and gene knockout studies. Bolouri described the importance of integrative analysis of large scale datasets for molecular biology and listed key findings of modeling and analyzing GRN [33]. Two of these findings constitute the imperative for approaches like ours:

- Current models of the regulation of gene expression are far too simplistic and need updating.
- Integrative computational analysis of large-scale datasets is becoming a fundamental component of molecular biology.

An example for data integration is the usage of gene expression experiments like microarray or RNA-seq, and DNA binding experiments like ChIP-seq to analyze regulatory mechanisms. Thereby gene expression data is used to extract similarly expressed TFs

from time-series experiments (during differentiation) and ChIP-seq experiments can indicate, whether co-localized binding sites exist for the co-expressed TFs. Another example for the benefit of data integration strategy is the combination of expression data and the verification of derived potential interactions using methods such as RNAi knockdown applied by Konig et al. to identify useful intervention strategies in infections [193].

In the context of T-cells, Ciofani et al. showed the impact of data integration for improving and verifying interactions in the gene regulatory network of T helper 17 (Th17) cell master TFs [66]. First, they derived potential TF – target gene interactions using TFBS detection and motif search based on Th17 cell ChIP-seq data. After establishing a baseline network, regulatory information for the integrated and additional TFs was added by integrating RNA-seq expression data from knock-out and wild-type mice and publicly available microarray expression experiments of other immune cells and conditions. The derived baseline network was enriched by a GRN that was inferred by analyzing the expression data with the Inferelator tool [36]. GSEA was used to characterize the predicted interactions and top-enriched genes were checked using literature search. Additionally, Ciofani et al. used GWAS to check associations between genes covered by the core network and Th17 functions as well as SNP of Th17 cell related diseases.

Here, we first describe our approach of integrating ChIP-seq, RNA-seq, microarray, literature and information from databases (Section 6.2). Second, in Section 6.3 we apply the presented approach to reconstruct a network for Th2 cells around its master TF Stat6. Aim of our data integration strategy was a better understanding of the T cell fate decisions of naïve Th cells during differentiation to Th2 cells.

6.2. Integrating ChIP-seq, RNA-seq, Microarrays data

To enable the integration of different experimental data, we developed an integration pipeline consisting of three parts (see Figure 6.1). After datasets selection (part 1), raw data must be processed and analyzed (part 2), e.g. differential expression analysis. The actual integration (part 3) is performed regarding necessary mapping between biological entities, as far as this has not be considered in previous steps.

Here, we present in detail the applied analysis pipelines and workflows for ChIP-seq, RNA-seq and microarray experiments. Afterwards, the integrations of the results is explained.

6.2.1. Processing raw data from ChIP-seq, RNA-seq and microarray experiments

Prior to the integration of different experimental data and external knowledge, preprocessing and mapping to a common Id system is required. Pre-processing is individual to the specific data type, i.e. RNA-seq data analysis differs from microarray analysis. But, for both an consistent identifier mapping strategy is crucial to ensure that integration of different data types refers to identical biological instances (e.g. by integrating

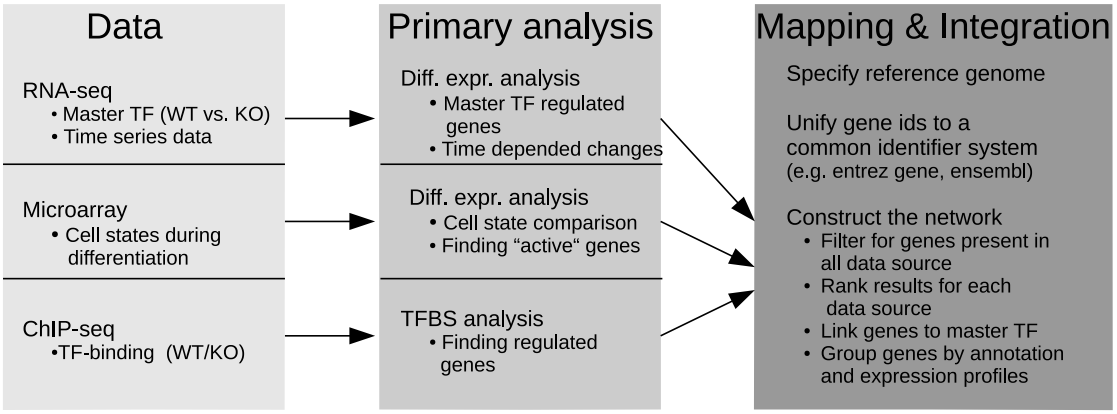


Figure 6.1.: **Integration strategy to create an interaction network around a master regulator (master TF).** This pipeline models the integration of four different data sources and is divided into three parts, i) data selection and preparation ii) analysis of each data set using specific workflows. iii) A mapping schema using a common id system is required to ensure compatibility between the data sets. Entities and links between the data sets are integrated into a single network based on unambiguous identifier only.

transcriptome, genome data). The next paragraph describes the processing workflows for the different data types we consider in this study.

ChIP-seq data analysis

For studying the genome-wide DNA-binding ChIP-seq is an appropriate method that enables the distinct analysis of TF binding and the analysis of the bound sequences and corresponding genes (see Section 2.4.2). ChIP-seq data analysis was performed based on the workflow in Figure 2.9 (p.19) and described as follows:

Raw read extraction Sequencing machine output (reads) conversion into fasta formatted files (fasta/fastq) using SRAtools software [322]

Quality control Read quality estimation using quality scores provided by the sequencing machine, statistics of multiple mappings, gc-content calculation and testing for nucleotide over-representation. FastQC [11] was used for quality control.

Trimming & filtering Read trimming is an essential step of cutting input read ends with poor quality values, as especially first generation sequencing machines tend to have a high error rate on read ends [372]. Using Solexa Trimming tool [74] reads were pre-processed to enable subsequent read mapping.

Read mapping Mapping the raw reads to a matching position in the prior selected reference genome using Bowtie2 [202] is a prerequisite for subsequent peak calling.

Peak calling After mapping the reads, peak calling is used to identify enriched genomic regions, i.e. regions with a high number of mapped reads. Next to high number of mapped reads also the shape of enrichment is accounted by the algorithms and

used to differentiate between random enrichment (false positives) and enrichment cause by TF-binding [150].

Peak annotation Associating peaks to genes enables the interpretation of the TF binding to specific genes. Search for the nearest transcription start side (TSS) of a gene around the peak is the straight forward way. While the average TFBS is located about 40bp up- or downstream of the TSS, other cis-regulatory elements are localized in greater distance to the TSS, e.g. enhancer, silencers, insulators and tethering elements [331]. We set the maximum allowed distance to 1.5kbp.

After finalizing the ChIP-seq analysis extracted peak annotations provide information about TF-binding in the promotor region of genes. This association between TF and genes can be integrated into a GRN as interaction between the TF and the bound genes.

RNA-seq data analysis

RNA-seq experiments are used to measure gene expression for the whole transcriptome. The advantage of RNA-seq for integration approaches is that RNA-seq enables the comparison of gene expression for different cells in different states (time points, treatment) between each other without restricting the analysis to prior defined regions or transcripts, as it is the case for microarray gene expression experiments (see next paragraph). To investigate differentiation processes of cells it is of high importance to acquire the cell state at different time-points and thus enable the comparison between the initial cell state (naive T cell) and the differentiation product (T helper cell subtype, like Th1, Th2).

The applied RNA-seq analysis pipeline is based on the workflow describe in Section 2.4.2 (Figure 2.10, p.20) and described as follows:

Raw read extraction & conversion Raw reads are retrieved and similar to ChIP-seq experiments files converted into fasta format files.

Quality control Calculating read statistics and checking for outlier is performed during quality control (see ChIP-seq analysis).

Read mapping We used Bowtie tool to perform the read alignment and TopHat2 tool for mapping to the selected reference genome [184]).

Transcript assembly We apply Cufflinks [359] to perform the assembly and merge of the reads to single genes or transcripts in the reference genome. The mapped reads are assembled and their abundances is estimated to test for differential expression and regulation in the RNA-Seq samples. Cufflinks estimates the relative abundances of transcripts based on how many reads support each one, taking library preparation protocols into account biases.

Differential expression analysis Cuffdiff [359] is used to compare gene/transcript expressions as well as identify differential expression.

Gene/transcript extraction Finally all raw measures, calculated expressions, fold changes and p-values are extracted.

Microarray data analysis

As gene expression data derived from RNA-seq, microarray gene expression data enables the detection of differences on transcriptome level (see workflow in Figure 2.6, p.16). Since the number of publicly available microarray experiments is still much higher than RNA-seq measurements such experiments are still very relevant for our integration approaches. One of the major drawbacks is that expression can be detected only for chip specific transcripts. After processing, as described in 2.4.1, we perform differential expression analysis using limma [329].

6.2.2. Data integration procedure

After processing and initial, data set specific analysis, the integration of individual results is performed (see Figure 6.1). Mapping to a common id system is the most important step of the integration.

We propose to ensure identifier compatibility already during the data analysis in part two of the experiment-type specific pipelines. For sequencing data analysis (ChIP-seq, RNA-seq) it is recommended to use identical reference genomes during read mapping and peak annotation to avoid ambiguity. For microarray analysis consistent mappings using a specific reference genome (e.g. the same as used for sequencing data) can be achieved by genome specific CDF files or mapping services like biomaRt [89]. The integration of different data sets is achieved by two steps:

1. Creating an adjacency matrix for each data set,
2. Combining adjacency matrices into a single network graph.

Adjacency matrices An adjacency matrix $C = N \cdot N$, where N is the number of genes and each entry in the matrix depicts the relation between two genes. In its simplest version the matrix is binary, where 1 represents an interaction and 0 no interaction, respectively. Adjacency matrices can also contain quantitative information, e.g. correlation values or probabilities of interactions. For example, differential expression analysis of KO and wild-type experiments (as used in Section 6.3). The expression change of genes can be used as quantitative information assigned to the relation of the measured gene and the knock-out gene. For each experiment an adjacency matrix is created by specified conditions, like TF binding for ChIP-seq, differential expression (RNA-seq KO experiments) or co-expression between two genes for microarray experiments are met. Table 6.1 provides an overview of the applied conditions.

Combining matrices The extracted adjacency matrices are incrementally added to the final network graph. After all integrated data sets are translated into binary adjacency matrices, the combination is performed by binary "OR" relation for all interactions. To conserve the information from particular not binary, adjacency matrices a multi-graph that enables parallel edges can be constructed. By adding parallel edges the individual properties and relation information are preserved.

Table 6.1.: **Summary of gene interactions derived from different experiments.** For knock-out (KO) experiments a relation is inferred between the knocked out gene and affected genes. For time-series or steady state experiments interactions are defined between co-regulated or co-expressed genes. (WT stands for wild-type)

Data	Sample classes	Derived gene-gene interactions, between
ChIP-seq	KO	KO-gene (TF) and bound genes (peak related)
RNA-seq	KO vs. wild-type (WT)	KO-gene and differentially expressed genes
	time series	co-regulated genes
	steady state	co-expressed genes, co-regulated genes
Mircroarray	KO vs. WT	KO-gene – differentially expressed genes
	time series	co-expressed genes
	steady state	co-expressed genes, co-regulated genes

Alternatively, the final network graph from all experiments can be iteratively constructed by adding an edge for each non-zero relation in the respective adjacency. This is done for each single adjacency matrix, i.e. for each experiment.

6.3. Reconstruction of a *Stat6*-centered network for Th2 cells

In this section, we describe in detail the integration strategy of a *Stat6* centered regulatory network for Th2 cells. The applied integration pipeline using ChIP-seq, RNA-seq, microarray, literature, and existing databases is illustrated in Figure 6.2. This pipeline led to a gene regulatory network around the TF STAT6, which fundamentally influences cell fate decision towards the Th2 cell subtype [390].

6.3.1. Preparation and analysis of publicly available data to reconstruct a gene regulatory network

We used selected public data sets to reconstruct gene regulatory networks, namely RNA-seq, microarray and ChIP-seq data sets. The initial network was established based on a *Stat6* KO mice RNA-seq experiment (GSE40463, [365]) and extended by publicly available RNA-seq and ChIP-seq experiments. In addition, AG Baumgrass (DRFZ) generated RNA-seq data of Th2 cell differentiation. GEO accession numbers and corresponding references are listed in Table 6.2 for all public data sets. We used published normalized *fragments per kilobase of exon per million fragments mapped (fpkm)* values to identify differentially expressed genes. Differential expression was defined as an absolute log_2(fold change) > 1. Low expressed genes were removed (normalized fpkm value smaller than 1 in all groups).

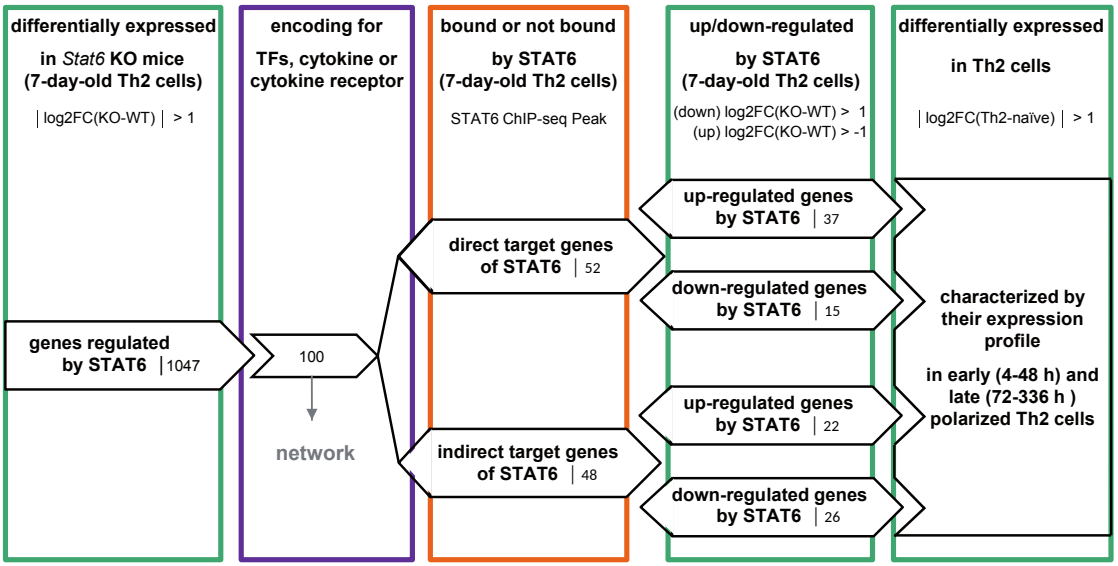


Figure 6.2.: **Workflow for the construction of the STAT6 network.** The used filter criteria and data sources for each step are shown in the first row and highlighted in different colors (green for RNA-seq, purple for microarray data, orange for ChIP-seq). The filter criteria and/or the number of genes are listed for each integration step and gene group in the second row. Figure taken from corresponding publication [168].

STAT6 deficient Th2 cells – ChIP-seq processing

We retrieved and integrated STAT6 ChIP-seq raw data of WT and STAT6-deficient Th2 cells by Wei et al. [379] (GSM550311, GSM550312 part of GSE22105). Reads were mapped to mouse genome using BWA (Burrows-Wheeler Aligner, [215]). To detect TF bindings, we applied CisGenome peak detection software to all uniquely mapped reads. As described by Wei et al., identical ChIP-seq peaks in KO and WT cells were removed prior to peak annotation. Genes associated to the peaks (by annotation) were integrated as binding targets respectively interaction partners of *Stat6*.

T helper cell time-series – RNA-seq data processing

Time-series expression experiment was generated using RNA-seq technology [168]. T cells were isolated, cultured and finally sequenced in an Ilumina sequencer HiSeq. After sequencing we performed a quality check and trimmed the sequences using Solexa trimming tool. The trimmed sequence reads where processed in exactly the same way as retrieved RNA-seq experiments by Hu et al. (GSE22081, [153]). All RNA-seq raw data was analyzed using the same pipeline to ensure comparability. Subsequently, raw reads were mapped and indexed using Tophat2 and Samtools. Read mapping was performed using Ensembl genome assembly mm10. Detected reads per gene (gene counts) were calculated using HTSeq. Next, differential expression was detected with R package DE-

Table 6.2.: **Integrated data sets for Stat6-Th2 cell network reconstruction.**

Data source	Cell type	Description	GEO accession number	Reference
RNA-seq	Th2 WT & Stat6 KO	naïve CD4+T cells cultured in vitro for 7 days under Th2 condition and restimulated with plate-bound CD3/28 +IL4 for 2 hrs	GSE40463	[365]
	Th1 WT & Stat4 KO	naïve CD4+T cells cultured in vitro for 7 days under Th1 condition and restimulated with plate-bound CD3/28 +IL12 for 2 hrs	GSE40463	[365]
	naive, Th1, Th2, Th17, iTreg	naïve CD4+T cells cultured in vitro for 4, 8, 12, 24, 48, 72, 168 and 336hrs under subtype specific conditions	GSE48138	[153]
	naive, Th1, Th2, iTreg	naïve CD4+T cells cultured in vitro for 5, 12, 24, 48 hrs) under subtype specific conditions	*DRFZ*	*unpublished*
Microarray	Th1, Th2, Th17, iTreg	naïve CD4+T cells cultured in vitro for 10 days under Th1, Th2, Th17 and iTreg conditions	GSE14308	[378]
ChIP-seq	Th2 WT & Stat6 KO	naïve CD4+T cells cultured in vitro for 7 days under Th2 condition and re-stimulated with plate-bound CD3/28 +IL4 for 2 hrs	GSE22105	[379]
	Th2 WT & Stat4 KO	naïve CD4+T cells cultured in vitro for 7 days under Th2 condition and restimulated with plate-bound CD3/28 +IL4 for 2 hrs	GSE22105	[379]

Seq. Thresholds for differentially expressed genes were defined by a minimal, absolute log_2 fold change of 1 and a p-value below 0.05. The resulting gene set is suspected to be highly influenced by *Stat6* KO and has been integrated as set of interaction partners of *Stat6* into the adjacency matrix.

Expression analysis of STAT6-regulated genes – microarray data processing

We characterized STAT6-regulated genes using expression levels of the integrated microarray data set (GSE14308, [378]) of the 10-day-old Th cells (Th2, Th1, Th17 and iTreg). We calculated the overlap of genes that were STAT6-regulated and significantly higher expressed after 10-days exclusively in one Th cell subtype compared to all others and thus we imply a certain Th subtype specificity after the differentiation (10 days, see Figure 6.5, p.104). For example, a gene expressed at least 1.4 fold higher in Th2 cells than in Th1, Th17 and iTreg cells indicated Th2 specificity. To illustrate this categories of T cell subtype-specificity was defined via comparing expression pattern, e.g. if the expression of a gene higher in Th2 than in Th1 cells (Th2 >Th1), and for the same gene it is high in Th2 than in Th17 cells (Th2 >Th17), and it is also higher Th2 than in iTreg cells (Th2 >iTreg). These categories were also used in Figure 6.5. The minimal thresholds for differential expression was set to log_2(fold change) = 0.5 ($\sim$ 1.4 fold).

6.3.2. Th2 cell specific gene regulatory network assembly

All preprocessed data sets were iteratively integrated into a adjacency matrix representing the final network. First, we identified STAT6-regulated genes by analyzing RNA-seq data of 7-day-old in vitro differentiated Th2 cells of WT and *Stat6* KO mice (GSE40463, highlighted in green in Figure 6.2). Gene expression analysis determined 1047 differentially expressed genes. These presumably STAT6-regulated genes were filtered for TFs, cytokines and cytokine receptors based on a published Th cell microarray data set (GSE14308, in purple in Figure 6.2) of 782 TFs and 271 cytokines/cytokine receptors in the context of induction, differentiation and maintenance of Th cell subsets.

From ChIP-seq analysis 4136 genes were considered as STAT6-bound (GSE22105, highlighted in orange in Figure 6.2). The overlap of the Stat6-bound genes and the set of TFs and cytokines/cytokine receptors revealed exactly 100 genes that were regulated also by Stat6 (from RNA-seq analysis). This gene set was further classified into two categories, either direct or indirect targets of STAT6 (using ChIP-seq binding data of 7-day-old Th2 cells).

Figure 6.3) illustrates the initial definition of the genes forming our basic STAT6 network.

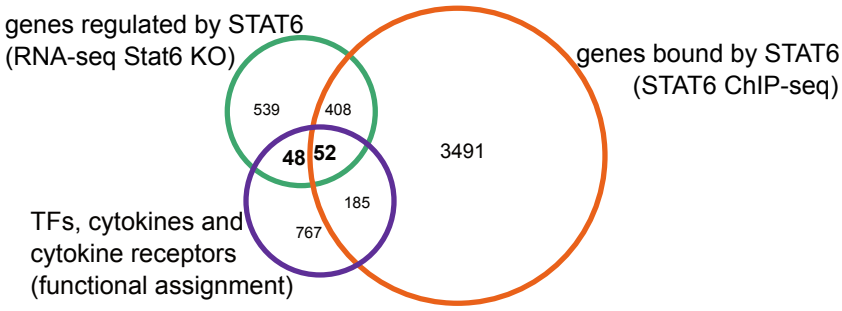


Figure 6.3.: **The Venn-diagram shows the three filtered gene sets forming the basic network.** The intersection (100 genes, highlighted bold) of STAT6-regulated genes (green) with TFs, cytokines and cytokine receptors (purple) are dissected into direct and indirect target genes of STAT6 by integrating the STAT6 ChIP-seq data (orange). Figure adapted from [168].

Next, we integrated both (GSE48138, DRFZ-RNA-seq data) time series expression data of Th2 cells to check, whether the genes are differentially expressed (compared to naïve CD4+ T cells) in early phase (0-48 h) and late phase (72-336 h) of Th2 cell differentiation. The estimation of times of high expression can help to characterize the specific role of an transcription factor or cytokine during the differentiation process. We characterized 14 to be early expressed, 14 late expressed, 70 early and late expressed genes as well as two genes neither differentially expressed early or late. In the next step, the genes were connected to STAT6 regarding their co-expression in the Th2-RNA-seq experiments. In the final network visualization (Figure 6.4) edge colors indicate whether gene expression was increased or decreased in *Stat6* KO mice. Furthermore, genes were

annotated with their expression in other Th cell subtypes based on RNA-seq data and experiments by Hu et al. (GSE48138, [153]).

The resulting STAT6 network contains the following information: (i) how STAT6 influences the expression of other genes, (ii) whether genes encode for TFs or cytokines/cytokine receptors, (iii) whether genes are directly bound or not by STAT6, and (iv) whether genes are differentially expressed in early and/or late differentiation of Th2 cells compared to naïve CD4+ T cells.

6.3.3. Implications from the reconstructed network

The resulting gene regulatory network includes 100 TFs, cytokines and cytokine receptors regulated (48 genes) or directly targeted by STAT6 (52 genes) in Th2 cells (Figure 6.4). Of these, 59 genes are up- and 41 genes are down-regulated. The integration of expression time series of Th2 cells revealed that 98 of these 100 genes are also differentially expressed in Th2 cells compared to naïve T cells. Most of them are differentially expressed in early and late phase of differentiation.

The network contains genes known to be active in Th2 cells, such as the master TF of Th2 cells, *Gata3*, the Th2 cell specific surface receptors *Ccr3*, *Ccr4*, and *Ccr8*, and also the cytokines *Il4* and *Il6*, which are known to be positively regulated by STAT6 [35, 78, 390]. Importantly, master TFs for other Th cell subsets are part of the network but negatively regulated by STAT6, such as *Foxp3* for iTreg cells, *Tbx21* for Th1 cells, and *Rorg* for in Th17 cells [390]. Th2 cell specific TFs are positively regulated by STAT6 and increased in Th2 cells such as *Gata3* and *Satb1*.

Furthermore, we filtered the genes for TFs and performed a literature search regarding their function in cell differentiation processes. Table 6.3 summarizes the results from the presented experiments and provides a short functional description and references to related literature. The total set of 32 TFs is divided into three groups: (i) associated with differentiation of Th2 cells, (ii) associated with differentiation of other Th cell subtypes and (iii) so far not reported to be associated with differentiation of Th cells. While the majority of STAT6-negatively regulated TFs are activators for Th1, Th17 or iTreg cell differentiation, five out of six Th2 cell specific TFs are up-regulated by STAT6 (see Tabel 6.3). In summary, STAT6-driven differentiation of Th2 cells is mainly modulated by activators and not preferentially by repressors.

6.3.4. Evaluation of the network

To validate the Th2 cell specificity of the STAT6 network, we studied gene expression profiles in Th cells of the 100 STAT6-regulated genes. To confirm the hypothesis that STAT6-positively regulated genes have a higher expression preference in Th2 cells compared to STAT6-negatively regulated genes we compared the gene expression levels for each of the 100 STAT6-regulated genes in 10-day-old in vitro differentiated Th1, Th2, Th17 and iTreg cells (GSE14308 [379]). There are 64 genes preferentially expressed in one Th-cell-subtype. We subdivided this group into Th1-, Th2-, Th17- and iTreg-specifically-expressed genes and discriminated between STAT6-positively- (35) and

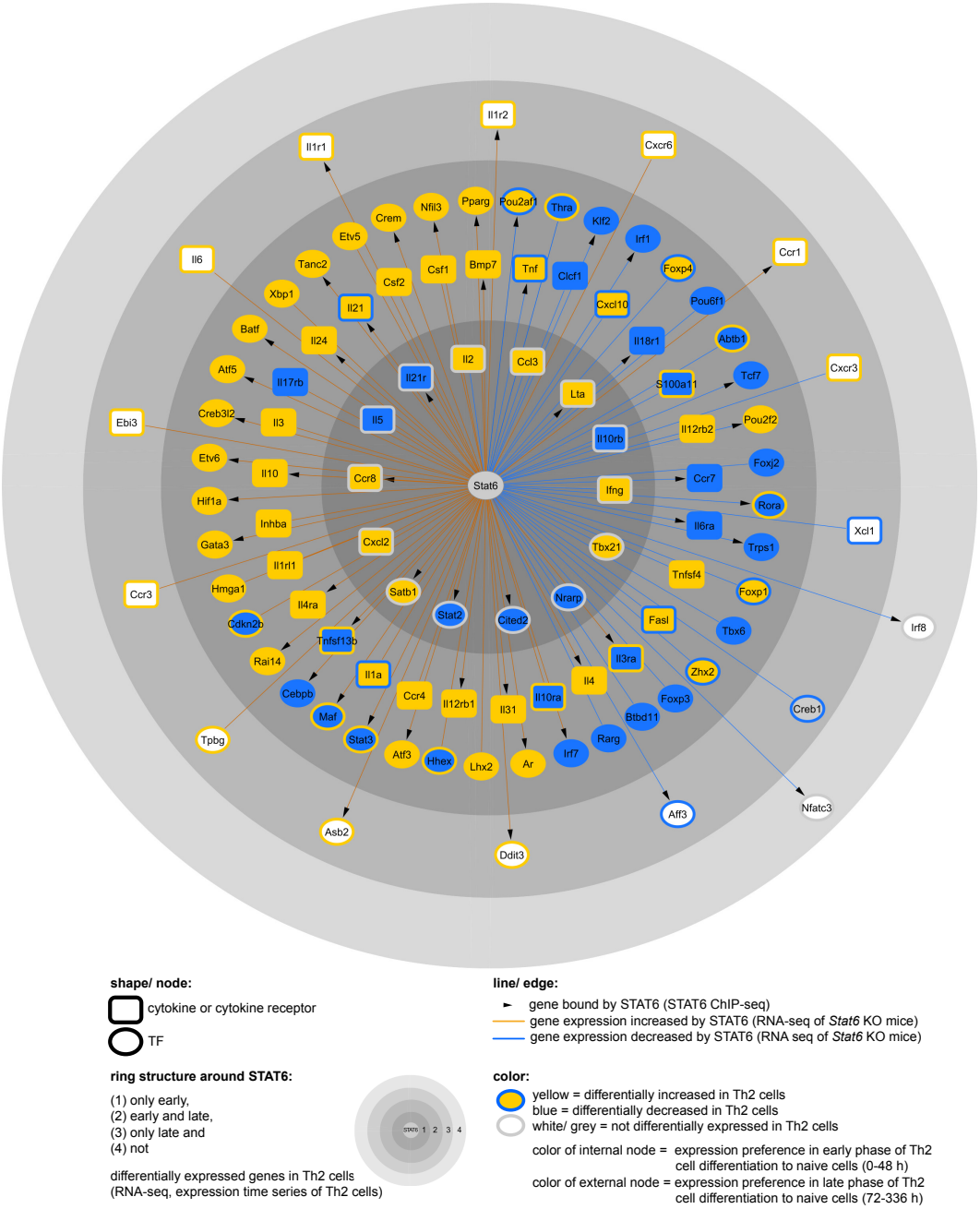


Figure 6.4.: **Gene regulatory network of STAT6-regulated genes in Th2 cells.** The network is limited to TFs (square node), cytokines or cytokine receptors (oval node) and shows indirect (48 genes) and direct targets (52 genes) of STAT6. The network layout is based on four rings around STAT6 in graded gray tones. Rings from inside to outside contain genes according to their expression behavior: genes are only early (1), early and late (2), only late (3) and not (4) differentially expressed in Th2 cells compared to naïve $CD4^+$ T cells. The edge color indicates the effect of *Stat6* on gene expression changes, while the node color indicates the expression changes compared to naïve $CD4^+$ T cells. Figure adapted from [168].

STAT6-negatively regulated (29) genes, as shown in the pie charts (Figure 6.5). The majority of the 35 STAT6-positively regulated genes (22) is preferentially expressed in Th2 cells. Importantly, there are no Th2-subtype-specific genes among the STAT6-negatively regulated genes validating the Th2 cell specificity of the STAT6 network. Furthermore, it confirms the observation that Th2-subtype-specific genes are up-regulated by STAT6 and Th1-, Th17- or iTreg-subtype-specific genes are mainly down-regulated by STAT6.

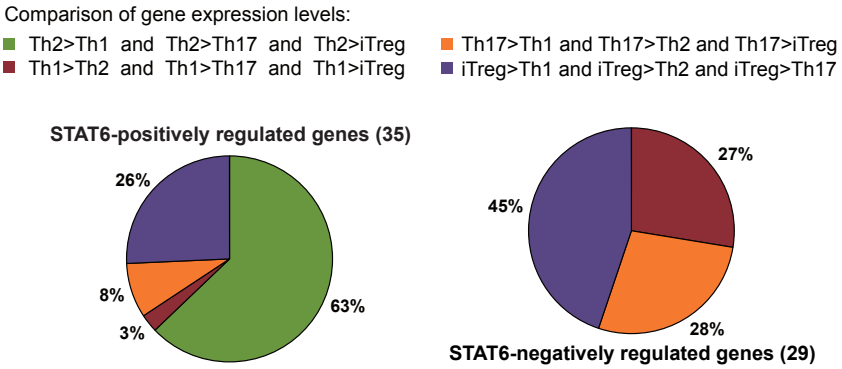


Figure 6.5.: **Expression preferences of Th cell subtype-specific and STAT6-regulated genes.** Gene expression levels of 100 STAT6-regulated genes were compared between the Th cell subtypes Th1, Th2, Th17, and iTreg. Figure adapted from [168].

6.3.5. Discussion of the presented integration strategy

In contrast to the large-scale network reconstruction approaches based only on expression data described in Chapter 5, we showed the power of integrating different, specific, and very homogenous experiments. The obvious advantage of this strategy is that data set size can be smaller and as a consequence of this, the impact of variability of the integrated data is smaller, too. Here, homogeneity depicts the experimental conditions for cell type distinction, culture methods, stimulation for T cell activation and intervals between measurements. Additionally, an integrated data normalization strategy covering all experiments at once, is not required because integration is performed at the results level, rather than on raw data level (in analogy to the meta-analysis strategy class described in Chapter 3). Also, the step-by-step integration of each single data set allows for adaption and changes of the analysis pipeline, without side effecting other parts. Furthermore, the pipeline is flexible towards improvements and extension in any of the three steps.

6.4. Summary

In summary, we applied comprehensive data integration to better understand Th2 cell fate decisions and, in particular, to delineate the gene regulatory network of Th2 cells to

unravel novel important TFs in Th2 cell differentiation. We identified 11 direct target genes of STAT6 with so far unreported functions in Th cell differentiation. Among these are eight genes positively (*Asb2, Atf5, Creb3l2, Cebpb, Cited2, Rai14, Tanc2*, and *Ddit3*) and three genes negatively regulated (*Aff3, Trps1*, and *Pou2f2*) by STAT6 (Table 6.3). Interestingly, two of these TFs (AFF3 and POU2F2) have not been yet directly associated to Th cell differentiation processes.

The eight STAT6-positively regulated and so far unknown TFs in the context of Th2 cell differentiation (Table 6.3) are of particular interest, because these TFs act as activators in Th2 or suppressors in Th1, Th17 or iTreg cell differentiation, as it was already shown for the other two groups.

Table 6.3.: **Functions in differentiation processes of STAT6-regulated and STAT6-bound genes encoding for TFs.** The 32 STAT6-regulated TFs were included into a literature search concerning the context of T cell differentiation within the publication period 2000-2015. Table is taken and adapted from [168].

STAT6-regulated and STAT6-bound genes encoding for TFs		Regulation of gene expression			Function described in literature	
		by Stat6	in polarized Th2 cells		for Th cell subtype	
			late	early		
Associated with differentiation of Th2 cells	Gata3	positive	↑	↑	Th2	master TF of Th2 cell differentiation [390]
	Atf3	positive	↑	↑	Th2	induced by IL4 and regulated by Stat6 [60]
	Satb1	positive	n.diff	↑	Th2	induced by IL-4 [3] and regulates Th2 cytokine gene expression [293]
	Nfil3	positive	↑	↑	Th2, Th17	regulates Th2 cytokine milieu: ↑ IL4, ↓ IL13 and IL5 [179] and inhibits Th17 cell differentiation [400]
	Maf	positive	↑	↓	Th2, Th17	positively regulates IL-4 expression [199] and induces Th17 cell differentiation via RORγt induction [349]
	Pparg	positive	↑	↑	Th2, Th1, Th17, iTreg	enhances production of Th2 and iTreg specific cytokines and decreases production of IFNγ and Th17 cell differentiation [77]
Associated with differentiation of other Th cell subtypes	Ar	positive	↑	↑	Th1	up-regulation of Ptpn1, which inhibits IL-12 signaling and Th1 cell differentiation [186]
	Irf7	positive	↓	↓	Th1	amplifies the anti-viral Th1 cell response by inducing the expression of genes encoding other type I IFNs [145, 235]
	Stat2	positive	n.diff	↓	Th1	mediates signaling by type I IFNs and induced Stat4 activation and Th1 cell development [100]
	Stat3	positive	↑	↓	Th2, Th17, iTreg	induces Th17 cell differentiation [406], promotes Th2 cell differentiation [340] and inhibits iTreg cell differentiation [204]
	Crem	positive	↑	↑	Th17	reduces activation of the AKT/mTOR pathway and enhance Th17 cell differentiation [190]
	Etv6	positive	↑	↑	Th17	repressor for Th17 cell differentiation [66]
	Hif1a	positive	↑	↑	Th17, iTreg	supports Th17 and inhibits iTreg cell differentiation through glycolytic shift [272]
	Batf	positive	↑	↑	Th17, Th9, Tfh	controls differentiation of Th17 [66, 314], Tfh [231] and Th9 cells [165]
	Irf1	negative	↓	↓	Th1, Th2	induces Th1 cell differentiation [403] and suppresses Il4 in response to IFNγ [93]
	Nfatc3	negative	n.diff	-	Th1, Th2	enhances Th1 cytokines IFNγ and TNFα, suppresses Th2 cytokines IL4 and IL5 in Th2 cells [60, 289]
	Tcf7	negative	↓	↓	Th1, Th2, Th17	induces Gata3 [399] and inhibits IFNγ and IL17 [240], downregulated by Il4/Stat6 and preferentially expressed in Th1 cells [238]
	Pou2af1	negative	↓	↑	Th1, Th17	regulates the balance of Th1 versus Th2 cell mediated immunity [43] and promotes Th17 cell differentiation [395]
	Irf8	negative	-	n.diff	Th1, Th17	promotes Th1 cell differentiation by modulating the function of APCs [403] and represses Th17 cell differentiation Ouyang2011
	Rora	negative	↑	↓	Th17	upregulated in Th17 cells [394]
	Klf2	negative	↓	↓	Th1, iTreg	increases IL2 [383] and CCR5 [294] expression and is important for iTreg cell differentiation [270]
Not yet reported to be associated with differentiation of Th cell subtypes	Asb2	positiv	↑	n.diff		inhibits growth and promotes commitment in myeloid leukemia cells (Guibal2002)
	Atf5	positiv	↑	↑		adipocyte differentiation (Zhao2014) and differentiation of oligodendrocytes (Mason2005)
	Cebpb	positiv	↓	↓		promotes differentiation of mammary epithelial cells [230], chondrocytes [364] and keratinocytes [299]
	Cited2	positiv	n.diff	↓		maintenance of hematopoietic homeostasis during embryogenesis [61]
	Creb312	positiv	↑	↑		important for differentiation of chondrocytes [162, 303]
	Ddit3	positiv	↑	n.diff		control of growth and differentiation of several cell lineages in inflammation and stress response [65]
	Rai14	positiv	↑	↑		porcine adipocyte differentiation [157]
	Tanc2	positiv	↑	↑		regulates dendritic spines and excitatory synapses [131]
	Aff3	negative	↓	n.diff		lymphoid development and oncogenesis [142, 348]
	Pou2f2	negative	↑	↑		B cell differentiation and mediates physical interaction with T cells [72, 178]
	Trps1	negative	↓	↓		differentiation and proliferation of chondrocyte [384]

7. Summary & Future work

7.1. Summary

In this thesis we studied how to apply machine learning methods for a better characterization of regulatory T cell subsets. In detail, we presented an improved strategy for the identification of Treg subtype specific surface markers and for the reconstruction of gene-regulatory mechanisms for Treg cells using publicly available data only. We collected hundreds of publicly available murine T cell specific microarray experiments, filtered and combined a substantial subset to a large microarray data set containing experiments from two Treg cell subtypes. We developed and evaluated a meta analysis strategy as baseline for the identification of marker candidates using an ensemble feature selection approach based on this data set. Furthermore, we used the same data to reconstruct gene regulatory networks. To this end, we applied a set of existing reconstruction algorithms and compared the inferred networks.

The meta analysis in **Chapter 3** describes a novel approach for aggregation of publicly available Treg cell specific gene expression experiments to form a large meta experiment for further use. After selecting and collecting hundreds of studies from a public repository, manual curation by a trained immunologist, and filtering was performed to ensure biological comparability. In the next step, quality checks based on downloaded raw data led to a final set of 154 microarray experiments. Subsequent meta analysis was implemented in two steps. First, experiments were normalized study-wisely, i.e. all experiments belonging to the same study were preprocessed and normalized. In the second step, quartile normalization was applied to the full set to adjust for inter-study effects (also called batch- or lab-effects). The final meta expression set consisted of 154 expression values for each of the 12825 extracted genes. 125 nTreg and 29 iTreg cell experiments built the final data set. For quality control we analyzed expression value distributions before and after re-normalization as well as by differential expression analysis. Subsequently we assessed GO annotation enrichment of the set, which showed that its set-characteristics, e.g. enriched molecular function terms of differentially expressed genes did not change.

In **Chapter 4** we developed a method to identify six Treg cell subtype specific surface marker candidates from the meta expression set. We compared multiple feature selection methods and applied an ensemble of two methods to detect cell type specific gene markers, followed by SVM-based evaluation of extracted candidate genes. We extracted 41 marker gene candidates. These genes and corresponding proteins were curated for surface marker properties, which led to a final set of six genes. Subsequently, these genes

were experimentally validated using qRT-PCR to verify gene expression differences between iTreg and nTreg cells as obtained from the meta expression data. For three of these, namely, *Ctla2b*, *Emp1* and *Tnfrsf13b* we could confirm higher expression in iTreg cells, while for *Gabrr2*, *Cd97* and *Cd79b* we confirmed the higher expression in nTreg cells. Additionally, GO term annotation indicated an enrichment of the 41 extracted genes towards cell cycle-related functions. Secondary biological analysis revealed that the differences between the compared Treg cell data sets potentially derived from the substantial differences of the cell culture conditions applied before performing microarray experiments. Nevertheless, at least one of the candidates (*Emp1*) remained valid as its expression profile was differential in qRT-PCR experiments under identical culture conditions for both Treg cell subtypes. In summary, the applicability and validity of the approach could be shown, as we were able to identify substantial differences between two groups. However, it remains unclear, whether these genes indicate genomic differences between the distinct Treg cell subtypes, or are artifacts caused by culture condition or in the end a mixture of both.

Chapter 5 compares GRN reconstruction methods for regulatory T cells. This comparison revealed large differences regarding the need for prior knowledge like regulatory interactions or pre-classification of genes into known regulators, e.g. TFs and targets of regulation. Another important difference is that some methods require reduction in data variability by discretization or binning of values to reduce calculation complexity. Finally, we applied 11 methods to the Treg cell meta expression set. We used two public interaction databases to evaluate the inferred networks and, hence, tool performances. We could extract a small set of overlapping edges between the inferred networks. Some of these edges are relevant for Treg cell specific regulatory mechanism, as they include highly specific genes like *Foxp3* – the master TF of Treg cells. Furthermore, two cytokines, namely *Il-2* and *Lta*, both relevant for Treg cell differentiation, were part of these subnetworks. However, the poor overal accuracy of all reconstruction performances illustrate the difficulty of the GRN inference problem based on expression data only. Low numbers of true positive interactions underline the need for improvements in future work.

In contrast to the results from Chapter 5, the data integration strategy presented in **Chapter 6** created a gene interaction network with substantial numbers of T cell relevant interactions. Here, we reconstructed a network around *Stat6* – the master TF of Th2 cells, that describes regulatory interactions occurring during cell differentiation from nTreg and iTreg cell experiments. The network was established based on STAT6-knockout RNA-seq and microarray data combined with STAT6 DNA-binding data from ChIP-seq experiments. It was based on TFs and cytokines relevant for T cell development, which were extracted from literature data. By using , The initial network was built using the expression pattern of this reduced set of potential interaction partners. Only those data from specific phases (time points) during the cell differentiation process from time-series RNA-seq data was used. The subsequent annotation of binding-site information (from ChIp-seq data) added a physical dimension to a subset of the ex-

tracted interactions. The derived (star) network contained 100 edges – one interaction between each gene and *Stat6*. We identified two groups of interaction partners of *Stat6*: (i) 52 genes which may directly interact with *Stat6* via binding of the TF and (ii) 48 genes which probably are only individually regulated during Th2 cell differentiation. Interestingly, 11 of these genes were so far unknown to be involved in Th2 cell differentiation. Our integrative analysis approach revealed that eight of these 11 TFs are either activators or suppressors in other T cell subtypes.

7.2. Future directions

7.2.1. Meta-analysis

The bases of all subsequent analyses is the underlying data. In Chapter 3, we described a strategy for selecting and curating experimental data to ensure biological comparability. We were faced with variations in the experimental setup and technology (platform) of some selected studies, that did not match and led to exclusions of the corresponding experiments. To reduce the impact of technical and processing variations towards later analysis, more homogeneous data are required, on one hand. While on the other hand we performed additional experiments to confirm our findings.

A great opportunity for the aggregation of large transcriptome data sets lies in the integration of high-throughput sequencing data (RNA-seq) to extend the amount of available experiments and studies. For about 10 years the amount of available biological sequence data has continuously increased continuously (see Figure 7.1). Dozens of studies assessed and confirmed the comparability of RNA-seq and microarray experiments (for a selection see [115, 245, 250, 284]). The main advantage of RNA-seq in contrast to microarray technology is its superiority in detecting low abundance transcripts, differentiating biologically critical isoforms, and allowing the identification of genetic variants [404]. Since RNA-seq experiments have certain advantages, we suggest to improve the presented meta analysis strategy by combining both technologies (microarray and RNA-seq). The combination of technologies enlarges the amount of available experiments and consequently the resulting expression set. Still, the challenge of handling study-effects exists for both technologies. Ma et al. analyzed the effect on RNA-seq data sets and observed study-effects to be consistent across microarray and RNA-seq data sets [241]. Therefore, the adaption of meta-analysis methods to RNA-seq data [251, 257] is needed. Also, it would be highly interesting to analyze the performance of the presented two-step normalization approach for reducing the batch effect. Note that, especially the second phase is not restricted to a specific platform or technology.

7.2.2. Cell specific surface marker identification

The presented approach of detecting surface marker candidates to enable distinction between Treg subtypes has not been tested *in vivo* so far. Consequently, the next step towards *in vivo* application is transferring the results from transcriptome to proteome level. This requires validation of candidates on protein level via immunoblot analysis

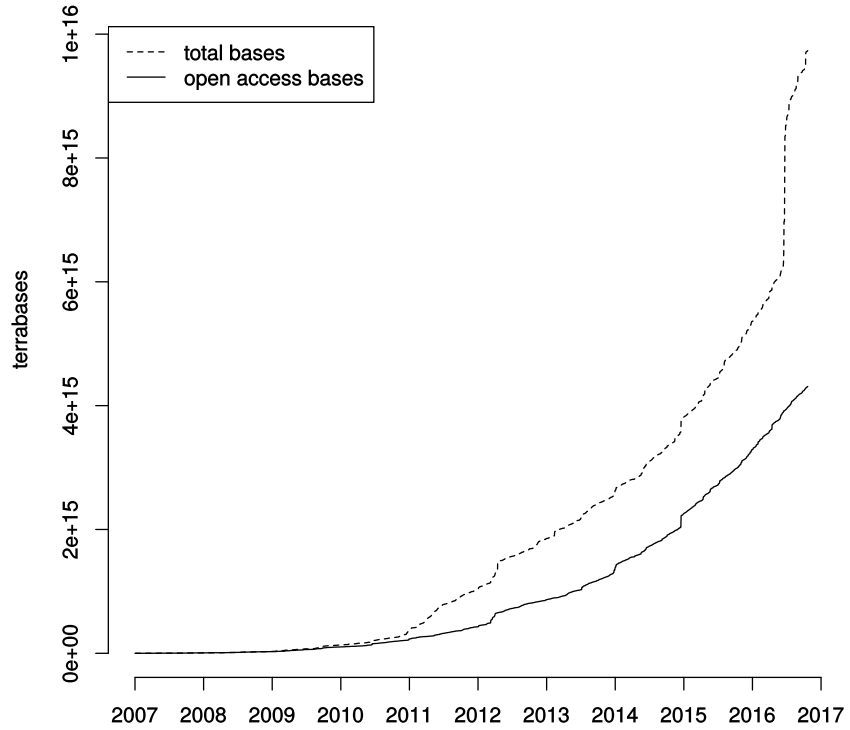


Figure 7.1.: **Number of bases uploaded to the Sequence Read Archive (SRA) [209].** The SRA together with the European Nucleotide Archive (ENA) [210] and the DNA Database of Japan [189] are the biggest open repository for sequence data. All three are partners in the International Nucleotide Sequence Database Collaboration (INSDC), that agreed on sharing all experiments.

(for details of the method see [106]), followed by flow cytometry based cell sorting. So far, testing of all extracted candidates using this approach was not possible due to a lack of working antibodies and the discussed shortcomings regarding culture conditions of the experiments.

Novel antibody panels (described in [118] or on manufacturer website[2]) covering hundreds of surface marker proteins, allow for the detection of cell type specific marker protein expression by flow cytometry. Such panels would allow for the parallel testing of multiple candidates and multiple cell types.

7.2.3. Gene regulatory network reconstruction

In this thesis, we described the reconstruction of Treg cell specific GRNs using a selection of methods based on steady-state data, which led to unsatisfying results. Quality of input data strongly influences GRN reconstruction results, similar to surface marker

[2]online: `http://www.biolegend.com/legendscreen`, August 1st, 2016

identification discussed in Chapter 4. As a consequence, we postulate also for GRN reconstruction more and better input data for future attempts.

We highlighted two other aspects for improving network inference. First, time-series data and second, more specific and smaller gene sets prior to inference. The benefit of time-series data has been tested for GRN reconstruction in various studies [27, 112, 315, 373], especially for the reconstruction of dynamic networks that reflect different temporal aspects of regulatory mechanisms. However, the usage of steady-state data was justified by the low amounts of publicly available time-series data, that could be applied to meta analysis and subsequent GRN reconstruction.

Specifying gene subsets is another option to improve GRN reconstruction. Categorization of genes into regulated and regulating genes, as well as sub-grouping of genes based on expression profiles are two approaches for subset definition. Such could replace the presented strategy, of selecting the highest differentially expressed genes over all experiments. One approach is to define groups of differentially expressed genes regarding correlated temporal phases (time points). This extend existing approaches where time-delays in gene expression are used to determine regulatory mechanisms [135, 148, 410]. Since it is widely accepted that genes can be grouped into regulators and regulated ones, it may be worth to utilize time-series data to extract the groups from consecutive time-points. Decomposing the set of input genes regarding their expression profiles reduces the complexity of the inference task. As matter of fact, the results of the presented study indicate that GRN reconstruction performs better for small networks/gene sets (see Figure 5.5, p. 88).

Additionally, the genes to be studied could be more specifically selected regarding known regulatory functions or functional annotations, like TFs or miRNAs. Such categorization can also be incorporated as activity list or initial network structure.

Combining data integration and GRN reconstruction

Another approach for improving GRN reconstruction is to integrate prior knowledge from databases and related experiments into the inference process. The basic idea is to combine the results from data integration (Chapter 6) as prior knowledge with methods for GRN reconstruction (Chapter 5).

More precisely, the presented integration approach for Th2 cells revealed 100 genes that interact with the TF *Stat6* and, thus, form a network with 101 interactions. These interactions can be passed to reconstruction tools, like birta, MNI or NCA (for details see Section 5.5). Such tools require initial TF-networks or so-called activity matrices encoding known interaction (see Table 5.3, 76) as initializing parameter sets. Bayesian network approaches like Banjo allow for initial network structures to set mandatory and "forbidden" edges.

Additionally, the analysis of differentially expressed genes from time-series RNA-seq data, helps to identify simultaneously regulated genes. Such co-regulated genes are more likely to share regulatory mechanisms and could be used to define subsets for inference for specific time-points or phase of cell differentiation as well as for treatment or phenotype studies. Another open question is to evaluate how GRN reconstruction methods like MNI

(an ODE-based method) can benefit from comparative studies revealing specific gene sets of interest, for example revealed by gene-knockout experiments or drug-treatment studies. The major advantage of specific (and often smaller) gene sets is, that the number of potential nodes in a network is reduced by magnitudes.

In summary, applying the integration strategy to retrieve a baseline set of gene-regulatory interactions is one future challenge for improving GRN reconstruction. However, uncovering regulatory mechanisms and even single interactions can help to elucidate the regulatory program behind T helper cell differentiation.

A. Appendix

A.1. References to microarray experiments integrated into meta-analysis

The appendix provides detailed information about the microarray gene expression experiments used to create the meta-expression set in Chapter 3. Table A.1 lists all microarrays included in the meta-analysis. The table provides detailed information on Treg cell subtype, cell origin (mouse strain), literature references to corresponding publications as well as all available GEO accessions (GDS, GSE, GSM). Table A.2 extends Table A.1 by details on referenced literature for each listed microarray experiments.

Table A.1.: **Listing of the Treg cell specific microarray experiments included in the meta expression set (see Chapter 3).** Categorization for Treg subtype was performed by biologist for each single microarray. All information were retrieved from NCBI's GEO Platform `http://www.ncbi.nlm.nih.gov/geo`. Full literature references are listed in Table A.2 using PMIDs as key. Table is adapted from [198, in preparation].

Treg cell subtype	Mouse strain	Title	GEO Experiment Accession No.	GEO Microarray Accession No.	GEO Platform Accession No.	PMID
iTreg	TCR-HA-transgenic mice	dtg CD25+ TR_1	GSE12506	GSM314229	GPL339	18056346
iTreg	TCR-HA-transgenic mice	dtg CD25+ TR_2	GSE12506	GSM314230	GPL339	18056346
iTreg	TCR-HA-transgenic mice	dtg CD25+ TR_3	GSE12506	GSM314231	GPL339	18056346
nTreg	C57BL/6-Foxp3/GFP	Foxp3-GFP mesenteric lymph node Treg, biological rep 1	GSE41229	GSM1011464	GPL6246	24574339
nTreg	C57BL/6-Foxp3/GFP	Foxp3-GFP mesenteric lymph node CD4+ non-Treg, biological rep 1	GSE41229	GSM1011465	GPL6246	24574339
nTreg	C57BL/6-Foxp3/GFP	Foxp3-GFP small intestine Treg, biological rep 1	GSE41229	GSM1011466	GPL6246	24574339
nTreg	C57BL/6-Foxp3/GFP	Foxp3-GFP small intestine Treg, biological rep 2	GSE41229	GSM1011467	GPL6246	24574339
nTreg	C57BL/6-Foxp3/GFP	Foxp3-GFP spleen Treg, biological rep 1	GSE41229	GSM1011480	GPL6246	24574339
nTreg	C57BL/6-Foxp3/GFP	Foxp3-GFP mesenteric lymph node Treg, bilogical rep 2	GSE41229	GSM1011481	GPL6246	24574339
nTreg	C57BL/6-Foxp3/GFP	Foxp3-GFP mesenteric lymph node Treg, bilogical rep 3	GSE41229	GSM1011493	GPL6246	24574339
nTreg	C57BL/6-Foxp3/GFP	Foxp3-GFP small intestine Treg, biological rep 3	GSE41229	GSM1011499	GPL6246	24574339
iTreg	C57BL/6J	Nr1+	GSE41492	GSM1018194	GPL1261	23226238
iTreg	C57BL/6J	Nr2+	GSE41492	GSM1018195	GPL1261	23226238
iTreg	C57BL/6J	Nr3-	GSE41492	GSM1018196	GPL1261	23226238
iTreg	C57BL/6J	Nr4-	GSE41492	GSM1018197	GPL1261	23226238
nTreg	Balbc	13_PTr	GSE42021	GSM1030709	GPL1261	23420886

Table A.1 – *Continued from previous page*

Treg cell subtype	Mouse strain	Title	GEO Experiment Accession No.	GEO Microarray Accession No.	GEO Platform Accession No.	PMID
nTreg	Balbc	14_PTr	GSE42021	GSM1030718	GPL1261	23420886
nTreg	Balbc	15_PTr	GSE42021	GSM1030727	GPL1261	23420886
nTreg	C57BL/6-Foxp3/GFP	TR.0hr#3	GSE42276	GSM1036805	GPL6246	23277554
nTreg	C57BL/6-Foxp3/GFP	TR.0hr#4	GSE42276	GSM1036806	GPL6246	23277554
nTreg	C57BL/6-Foxp3/GFP	TR.0hr#5	GSE42276	GSM1036807	GPL6246	23277554
nTreg	C57BL/6-Foxp3/GFP	Foxp3+ ex vivo_Cytoplasmic RNA_1	GSE45401	GSM1103573	GPL7546	23658533
nTreg	C57BL/6-Foxp3/GFP	Foxp3+ ex vivo_Polysome-associated RNA_2	GSE45401	GSM1103574	GPL7546	23658533
nTreg	C57BL/6-Foxp3/GFP	Foxp3+ ex vivo_Cytoplasmic RNA_5	GSE45401	GSM1103577	GPL7546	23658533
nTreg	C57BL/6-Foxp3/GFP	Foxp3+ ex vivo_Polysome-associated RNA_6	GSE45401	GSM1103578	GPL7546	23658533
nTreg	C57BL/6-Foxp3/GFP	Foxp3+ activated in vitro_Cytoplasmic RNA_13	GSE45401	GSM1103585	GPL7546	23658533
nTreg	C57BL/6-Foxp3/GFP	Foxp3+ activated in vitro_Polysome-associated RNA_14	GSE45401	GSM1103586	GPL7546	23658533
nTreg	C57BL/6-Foxp3/GFP	Foxp3+ activated in vitro_Cytoplasmic RNA_15	GSE45401	GSM1103587	GPL7546	23658533
nTreg	C57BL/6-Foxp3/GFP	Foxp3+ activated in vitro_Polysome-associated RNA_16	GSE45401	GSM1103588	GPL7546	23658533
nTreg	C57BL/6J	Treg_WT_replicate_1	GSE46693	GSM1134630	GPL11180	23812589
nTreg	C57BL/6J	Treg_WT_replicate_2	GSE46693	GSM1134631	GPL11180	23812589
nTreg	C57BL/6J	Treg_WT_replicate_3	GSE46693	GSM1134634	GPL11180	23812589
nTreg	C57BL/6J	Treg_WT_replicate_4	GSE46693	GSM1134636	GPL11180	23812589
nTreg	NOD	TR.VIVO.0hr#1	GSE48210	GSM1172803	GPL6246	23986534
nTreg	NOD	TR.VIVO.0hr#2	GSE48210	GSM1172814	GPL6246	23986534
nTreg	NOD	TR.VITRO.0hr#1	GSE48210	GSM1172825	GPL6246	23986534
nTreg	NOD	TR.VITRO.0hr#2	GSE48210	GSM1172835	GPL6246	23986534
nTreg	C57BL/6J	CD4+CD25+ Treg, B6.H2g7, rep1	GSE6813	GSM155637	GPL1261	19073938
nTreg	C57BL/6J	CD4+CD25+ Treg, B6.H2g7, rep2	GSE6813	GSM155638	GPL1261	19073938
nTreg	C57BL/6J	CD4+CD25+ Treg, B6.H2g7, rep3	GSE6813	GSM155639	GPL1261	19073938
nTreg	NOD	CD4+CD25+ Treg, NOD, rep1	GSE6813	GSM155640	GPL1261	19073938
nTreg	NOD	CD4+CD25+ Treg, NOD, rep2	GSE6813	GSM155641	GPL1261	19073938
nTreg	NOD	CD4+CD25+ Treg, NOD, rep3	GSE6813	GSM155642	GPL1261	19073938
nTreg	C57BL/6-Foxp3/GFP	76091_CD4+, FoxP3(EGFP+) mouse cells	GSE6875	GSM158520	GPL1261	17273171
nTreg	C57BL/6-Foxp3/GFP	82428_CD4+, FoxP3(EGFP+) mouse cells.txt	GSE6875	GSM158523	GPL1261	17273171
nTreg	C57BL/6-Foxp3/GFP	85382_CD4+, FoxP3(EGFP+) mouse cells	GSE6875	GSM158527	GPL1261	17273171
iTreg	NOD	ActTreg, rep1	GSE7460	GSM176764	GPL1261	18024188
iTreg	NOD	ActTregTGF, rep1	GSE7460	GSM176765	GPL1261	18024188
iTreg	NOD	ActTreg, rep2	GSE7460	GSM177375	GPL1261	18024188
iTreg	NOD	ActTreg, rep3	GSE7460	GSM177376	GPL1261	18024188
iTreg	NOD	ActTregTGF, rep2	GSE7460	GSM177377	GPL1261	18024188
iTreg	NOD	ActTregTGF, rep3	GSE7460	GSM177378	GPL1261	18024188
nTreg	C57BL/6J	Foxp3 sufficient Treg CD4+ T cells from healthy B6 mice_1	GSE11775	GSM298104	GPL1261	19710455, 21642545
nTreg	C57BL/6J	Foxp3 sufficient Treg CD4+ T cells from healthy B6 mice_2	GSE11775	GSM298105	GPL1261	19710455, 21642545
nTreg	C57BL/6J	Foxp3 sufficient Treg CD4+ T cells from healthy B6 mice_3	GSE11775	GSM298106	GPL1261	19710455, 21642545

Table A.1 – *Continued from previous page*

Treg cell sub-type	Mouse strain	Title	GEO Experiment Accession No.	GEO Microarray Accession No.	GEO Platform Accession No.	PMID
nTreg	TCR-HA-transgenic mice	stg TR_1	GSE12506	GSM314235	GPL339	18056346
nTreg	TCR-HA-transgenic mice	stg TR_2	GSE12506	GSM314236	GPL339	18056346
nTreg	TCR-HA-transgenic mice	stg TR_3	GSE12506	GSM314237	GPL339	18056346
nTreg	Balb/c	WT TR_1	GSE12506	GSM314238	GPL339	18056346
nTreg	Balb/c	WT TR_2	GSE12506	GSM314239	GPL339	18056346
nTreg	Balb/c	WT TR_3	GSE12506	GSM314240	GPL339	18056346
iTreg	TCR-HA-transgenic mice	stg 3d TA_1	GSE12506	GSM314241	GPL339	18056346
iTreg	TCR-HA-transgenic mice	stg 3d TA_2	GSE12506	GSM314242	GPL339	18056346
iTreg	TCR-HA-transgenic mice	stg 3d TA_3	GSE12506	GSM314243	GPL339	18056346
iTreg	C57BL/6J	iTreg-1	GSE14308	GSM357849	GPL1261	19144320
iTreg	C57BL/6J	iTreg-2	GSE14308	GSM357850	GPL1261	19144320
nTreg	C57BL/6J	nTreg-1	GSE14308	GSM357852	GPL1261	19144320
nTreg	C57BL/6J	nTreg-2	GSE14308	GSM357853	GPL1261	19144320
nTreg	C57BL/6J	C57BL/6 CD4+ CD25+ Treg cells #1	GSE14350	GSM358746	GPL1261	19185518
nTreg	C57BL/6J	C57BL/6 CD4+ CD25+ Treg cells #2	GSE14350	GSM358748	GPL1261	19185518
nTreg	C57BL/6J	C57BL/6 CD4+ CD25+ Treg cells #3	GSE14350	GSM358749	GPL1261	19185518
iTreg	Balbc/Foxp3EGFP	Induced regulatory T cells 1	GSE14415	GSM360147	GPL1261	19265124
iTreg	Balbc/Foxp3EGFP	Induced regulatory T cells 2	GSE14415	GSM360148	GPL1261	19265124
iTreg	Balbc/Foxp3EGFP	Induced regulatory T cells 3	GSE14415	GSM360149	GPL1261	19265124
iTreg	Balbc/Foxp3EGFP	Induced regulatory T cells 4	GSE14415	GSM360150	GPL1261	19265124
iTreg	Balbc/Foxp3EGFP	Induced regulatory T cells 5	GSE14415	GSM360151	GPL1261	19265124
nTreg	Balbc/Foxp3EGFP	Activated natural regulatory T cells 1	GSE14415	GSM360162	GPL1261	19265124
nTreg	Balbc/Foxp3EGFP	Activated natural regulatory T cells 2	GSE14415	GSM360163	GPL1261	19265124
nTreg	Balbc/Foxp3EGFP	Activated natural regulatory T cells 3	GSE14415	GSM360164	GPL1261	19265124
nTreg	Balbc/Foxp3EGFP	Native (not activated) natural regulatory T cells 1	GSE14415	GSM360165	GPL1261	19265124
nTreg	Balbc/Foxp3EGFP	Native (not activated) natural regulatory T cells 2	GSE14415	GSM360166	GPL1261	19265124
nTreg	Balbc/Foxp3EGFP	Native (not activated) natural regulatory T cells 3	GSE14415	GSM360167	GPL1261	19265124
nTreg	Balbc/Foxp3EGFP	Native (not activated) natural regulatory T cells 4	GSE14415	GSM360168	GPL1261	19265124
iTreg	Balbc/Foxp3EGFP	IL-2-cultured induced regulatory T cells 1	GSE14415	GSM360174	GPL1261	19265124
iTreg	Balbc/Foxp3EGFP	IL-2-cultured induced regulatory T cells 2	GSE14415	GSM360175	GPL1261	19265124
nTreg	129/SvJ/Foxp3-IRES-GFP	Naïve Treg_biological_rep4	GSE16210	GSM407042	GPL1261	Bioproject PRJNA117149
nTreg	129/SvJ/Foxp3-IRES-GFP	Naïve Treg_biological_rep5	GSE16210	GSM407043	GPL1261	Bioproject PRJNA117149
nTreg	129/SvJ/Foxp3-IRES-GFP	Naïve Treg_biological_rep6	GSE16210	GSM407044	GPL1261	Bioproject PRJNA117149
nTreg	C57BL/6J	Control Treg_rep1	GSE18148	GSM453647	GPL1261	19800266
nTreg	C57BL/6J	Control Treg_rep2	GSE18148	GSM453648	GPL1261	19800266
nTreg	C57BL/6J	Control Treg_rep3	GSE18148	GSM453649	GPL1261	19800266
nTreg	DEREG	CD4+GFP+FoxP3+	GSE18387	GSM458592	GPL1261	19966212
nTreg	C57BL/6-Foxp3/GFP	TregLPno1	GSE20366	GSM510238	GPL1261	20231436
nTreg	C57BL/6-Foxp3/GFP	TregLPno2	GSE20366	GSM510239	GPL1261	20231436
nTreg	C57BL/6-Foxp3/GFP	TregCD103+Klrg1+no1	GSE20366	GSM510252	GPL1261	20231436
nTreg	C57BL/6-Foxp3/GFP	TregCD103+Klrg1+no2	GSE20366	GSM510253	GPL1261	20231436
nTreg	C57BL/6-Foxp3/GFP	T reg resting replicate 1	GSE24210	GSM595507	GPL1261	20953201
nTreg	C57BL/6-Foxp3/GFP	T reg resting replicate 2	GSE24210	GSM595508	GPL1261	20953201

Table A.1 – *Continued from previous page*

Treg cell subtype	Mouse strain	Title	GEO Experiment Accession No.	GEO Microarray Accession No.	GEO Platform Accession No.	PMID
nTreg	C57BL/6-Foxp3/GFP	T reg resting replicate 3	GSE24210	GSM595509	GPL1261	20953201
nTreg	C57BL/6J	WT_1	GSE26425	GSM648480	GPL8321	21199917, 22715468
nTreg	C57BL/6J	WT_2	GSE26425	GSM648481	GPL8321	21199917, 22715468
nTreg	C57BL/6J	WT_3	GSE26425	GSM648482	GPL8321	21199917, 22715468
nTreg	C57BL/6J	WT 1	GSE27434	GSM678219	GPL8321	23444399
nTreg	C57BL/6J	WT 2	GSE27434	GSM678220	GPL8321	23444399
nTreg	C57BL/6J	WT 3	GSE27434	GSM678221	GPL8321	23444399
nTreg	C57BL/6J	WT_1	GSE27896	GSM688866	GPL8321	21444725
nTreg	C57BL/6J	WT_2	GSE27896	GSM688867	GPL8321	21444725
nTreg	C57BL/6J	WT_3	GSE27896	GSM688868	GPL8321	21444725
nTreg	C57BL/6-Foxp3/GFP	activated Treg, biological replica 1	GSE28130	GSM696984	GPL1261	21642545
nTreg	C57BL/6-Foxp3/GFP	activated Treg, biological replica 2	GSE28130	GSM696985	GPL1261	21642545
iTreg	C57BL/6-Foxp3/GFP	induced Treg, biological replica 1	GSE28130	GSM696986	GPL1261	21642545
iTreg	C57BL/6-Foxp3/GFP	induced Treg, biological replica 2	GSE28130	GSM696987	GPL1261	21642545
nTreg	C57BL/6J	Wild type Treg rep1	GSE29262	GSM723290	GPL1261	22013112
nTreg	C57BL/6J	Wild type Treg rep2	GSE29262	GSM723294	GPL1261	22013112
nTreg	C57BL/6J	Wild type Treg rep3	GSE29262	GSM723298	GPL1261	22013112
nTreg	NOD	CD25+CD49f+ T cells, biological rep1	GSE30503	GSM756576	GPL1261	Bioproject PRJNA143315
nTreg	NOD	CD25+CD49f+ T cells, biological rep2	GSE30503	GSM756577	GPL1261	Bioproject PRJNA143315
nTreg	NOD	CD25+CD49f+ T cells, biological rep3	GSE30503	GSM756578	GPL1261	Bioproject PRJNA143315
nTreg	NOD	CD25+CD49f+ T cells, biological rep4	GSE30503	GSM756579	GPL1261	Bioproject PRJNA143315
nTreg	NOD	NOD CRE natural Treg_MLB 004	GSE35164	GSM862871	GPL11180	22579476
nTreg	NOD	NOD CRE natural Treg_MLB 005	GSE35164	GSM862872	GPL11180	22579476
nTreg	NOD	NOD CRE natural Treg_MLB 006	GSE35164	GSM862873	GPL11180	22579476
nTreg	Balbc	Mouse invivo nTreg cells, replicate pool 1	GSE35543	GSM870351	GPL1261	23125413
nTreg	Balbc	Mouse invivo nTreg cells, replicate pool 2	GSE35543	GSM870352	GPL1261	23125413
iTreg	Balbc	Mouse stable invitro iTreg cells, replicate pool 1	GSE35543	GSM870353	GPL1261	23125413
iTreg	Balbc	Mouse stable invitro iTreg cells, replicate pool 2	GSE35543	GSM870354	GPL1261	23125413
nTreg	C57BL/6J	Wild type_1	GSE36095	GSM881027	GPL8321	22715468
nTreg	C57BL/6J	Wild type_2	GSE36095	GSM881028	GPL8321	22715468
nTreg	C57BL/6J	Wild type_3	GSE36095	GSM881029	GPL8321	22715468
nTreg	C57BL/6-Foxp3/RFP	FR1 (CD62Lhi CD69-Klrg1-) biological replicate 1	GSE36527	GSM896172	GPL6246	22786769
nTreg	C57BL/6-Foxp3/RFP	FR1 (CD62Lhi CD69-Klrg1-) biological replicate 2	GSE36527	GSM896173	GPL6246	22786769
nTreg	C57BL/6-Foxp3/RFP	FR2 (CD62Llo CD69-Klrg1-) biological replicate 1	GSE36527	GSM896174	GPL6246	22786769
nTreg	C57BL/6-Foxp3/RFP	FR2 (CD62Llo CD69-Klrg1-) biological replicate 2	GSE36527	GSM896175	GPL6246	22786769
nTreg	C57BL/6-Foxp3/RFP	FR3 (CD62Llo CD69+ Klrg1-) biological replicate 1	GSE36527	GSM896176	GPL6246	22786769
nTreg	C57BL/6-Foxp3/RFP	FR3 (CD62Llo CD69+ Klrg1-) biological replicate 2	GSE36527	GSM896177	GPL6246	22786769
nTreg	C57BL/6-Foxp3/RFP	FR3 (CD62Llo CD69+ Klrg1-) biological replicate 3	GSE36527	GSM896178	GPL6246	22786769

Table A.1 – *Continued from previous page*

Treg cell subtype	**Mouse strain**	**Title**	**GEO Experiment Accession No.**	**GEO Microarray Accession No.**	**GEO Platform Accession No.**	**PMID**
nTreg	C57BL/6-Foxp3/RFP	FR3 (CD62Llo CD69+ Klrg1-) biological replicate 4	GSE36527	GSM896179	GPL6246	22786769
nTreg	C57BL/6-Foxp3/RFP	FR3 (CD62Llo CD69+ Klrg1+) biological replicate 1	GSE36527	GSM896180	GPL6246	22786769
nTreg	C57BL/6-Foxp3/RFP	FR3 (CD62Lhi CD69- Klrg1-) biological replicate 10	GSE36527	GSM896181	GPL6246	22786769
nTreg	C57BL/6-Foxp3/GFP	TRGFP B6 #1	GSE37605	GSM923131	GPL6246	22579475
nTreg	C57BL/6-Foxp3/GFP	TRGFP B6 #2	GSE37605	GSM923132	GPL6246	22579475
nTreg	NOD/6-Foxp3/GFP	TRGFP NOD #1	GSE37605	GSM923133	GPL6246	22579475
nTreg	NOD/6-Foxp3/GFP	TRGFP NOD #2	GSE37605	GSM923134	GPL6246	22579475
nTreg	C57BL/6-Foxp3/IRES-GFP	TRIGFP B6 #1	GSE37605	GSM923135	GPL6246	22579475
nTreg	C57BL/6-Foxp3/IRES-GFP	TRIGFP B6 #2	GSE37605	GSM923136	GPL6246	22579475
nTreg	NOD/6-Foxp3/IRES-GFP	TRIGFP NOD #1	GSE37605	GSM923137	GPL6246	22579475
nTreg	NOD/6-Foxp3/IRES-GFP	TRIGFP NOD #2	GSE37605	GSM923138	GPL6246	22579475
nTreg	C57BL/6J	WT_1	GSE39864	GSM980505	GPL8321	22922362
nTreg	C57BL/6J	WT_2	GSE39864	GSM980506	GPL8321	22922362
nTreg	C57BL/6J	WT_3	GSE39864	GSM980507	GPL8321	22922362
nTreg	C57BL/6-Foxp3/GFP	WT 1 sample A	GSE40493	GSM995225	GPL6246	23053511
nTreg	C57BL/6-Foxp3/GFP	WT 1 sample B	GSE40493	GSM995226	GPL6246	23053511
nTreg	C57BL/6-Foxp3/GFP	WT 2 sample A	GSE40493	GSM995227	GPL6246	23053511
nTreg	C57BL/6-Foxp3/GFP	WT 2 sample B	GSE40493	GSM995228	GPL6246	23053511

Table A.2.: **Literature references of PMIDs corresponding to the experiments used to create the meta-expression set in Chapter 3.** The table is adapted from [198, in preparation].

PMID	Reference
17273171	Lin W, Haribhai D, Relland LM, et al.;Regulatory T cell development in the absence of functional Foxp3; Nat Immunol;2007 Apr;8(4):359-68;Epub 2007.
18024188	Hill JA, Feuerer M, Tash K, et al.; Foxp3 transcription-factor-dependent and -independent regulation of the regulatory T cell transcriptional signature; Immunity; 2007Nov;27(5):786-800.
18056346	Hansen W, Westendorf AM, Reinwald S,et al.; Chronic antigen stimulation in vivo induces a distinct population of antigen-specific Foxp3 CD25 regulatory T cells; J Immunol; 2007 Dec 15;179(12):8059-68;Erratum in: J Immunol. 2008 Oct15.181(8):5803.
19073938	D'Alise AM, Auyeung V, Feuerer M, et al.; The defect in T-cell regulation in NOD mice is an effect on the T-cell effectors;Proc Natl Acad Sci U S A; 2008 Dec
19144320	Wei G, Wei L, Zhu J, Zang C, et al.; Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells; Immunity; 2009.
19185518	Yu A, Zhu L, Altman NH, Malek TR; A low interleukin-2 receptor signaling threshold supports the development and homeostasis of T regulatory cells;Immunity; 2009.
19265124	Haribhai D, Lin W, Edwards B, et al.; A central role for induced regulatory T cells in tolerance induction in experimental colitis; J Immunol; 2009.
19710455	Kuczma M, Podolsky R, Garge N, et al.;Foxp3-deficient regulatory T cells do not revert into conventional effector CD4+ T cells but constitute a unique cell subset; J Immunol; 2009.
19800266	Kitoh A, Ono M, Naoe Y, et al.; Indispensable role of the Runx1-Cbfbeta transcription complex for in vivo-suppressive function of FoxP3+ regulatory T cells; Immunity; 2009.
19966212	Anz D, Koelzer VH, Moder S, et al.;Immunostimulatory RNA blocks suppression by regulatory T cells; J Immunol; 2010.
20231436	Feuerer M, Hill JA, Kretschmer K, et al.; Genomic definition of multiple ex vivo regulatory T cell subphenotypes; Proc Natl AcadSci U S A; 2010.
20953201	Collison LW, Chaturvedi V, Henderson AL, et al.; IL-35-mediated induction of a potent regulatory T cell population; Nat Immunol; 2010.
21199917	Beier UH, Wang L, Bhatti TR, et al.; Sirtuin-1targeting promotes Foxp3+ T-regulatory cell function and prolongs allograft survival; Mol Cell Biol; 2011..
21444725	de Zoeten EF, Wang L, Butler K, et al.; Histone deacetylase 6 andheat shock protein 90 control the functions of Foxp3(+) T-regulatory cells; MolCell Biol; 201.
21642545	Kuczma M, Lee JR, Kraj P; Connexin 43 signaling enhances the generation ofFoxp3+ regulatory T cells; J Immunol; 2011.
22013112	Pillai MR, Collison LW, Wang X, et al.; The plasticity of regulatory T cell function; J Immunol; 2011.
22579475	Darce J, Rudra D, Li L, et al.; An N-terminal mutation of the Foxp3 transcription factor alleviates arthritis but exacerbates diabetes; Immunity; 2012.
22579476	Bettini ML, Pan F, Bettini M, et al.; Loss of epigenetic modification driven by the Foxp3 transcription factor leads to regulatory T cell in sufficiency; Immunity; 2012.

Table A.2 – *Continued from previous page*

PMID	Reference
22715468	Beier UH, Wang L, Han R, et al.; Histone deacetylases 6and 9 and sirtuin-1 control Foxp3+ regulatory T cell function through shared and isoform-specific mechanisms; Sci Signal; 2012.
22786769	Cheng G, Yuan X, Tsai MS, et al.; IL-2 receptor signaling is essential for the development of Klrg1+ terminally differentiated T regulatory cells; J Immunol; 2012.
22922362	Rudra D, deRoos P, Chaudhry A, et al.; Transcription factor Foxp3 and its protein partners form a complex regulatory network; Nat Immunol; 2012.
23053511	Sawant DV, Sehra S, Nguyen ET, et al.; Bcl6 controls the Th2 inflammatory activity of regulatory T cells by repressing Gata3 function; JImmunol; 2012.
23125413	Schmitt EG, Haribhai D, Williams JB, et al.; IL-10 produced by induced regulatory T cells(iTregs) controls colitis and pathogenic ex-iTregs during immunotherapy; JImmunol; 2012.
23226238	Langenhorst D, Gogishvili T, Ribechini E, et al.; Sequential induction of effector function, tissue migration and cell death during polyclonal activation of mouse regulatory T-cells; PLoS One;2012.
23277554	Wakamatsu E, Mathis D, Benoist C; Convergent and divergent effects of costimulatory molecules in conventional and regulatory CD4+ T cells; Proc NatlAcad Sci U S A; 2013.
23420886	Toker A, Engelbert D, Garg G, et al.; Active demethylation of the Foxp3 locus leads to the generation of stable regulatory T cells within the thymus; J Immunol; 2013.
23444399	Wang L, Liu Y, Beier UH, et al.; Foxp3+T-regulatory cells require DNA methyl transferase 1 expression to preven tdevelopment of lethal autoimmunity; Blood; 2013.
23658533	Bjur E, Larsson O, Yurchenko E, et al.; Distinct translational control in CD4+ T cell subsets; PLoS Genet; 2013.
23812589	Zeng H, Yang K, Cloer C, et al.; mTORC1 couples immune signals and metabolic programming to establish T(reg)-cell function; Nature; 2013.
23986534	Li L, Nishio J, van Maurik A, Mathis D, et al.; Differential response of regulatory and conventional CD4+ lymphocytes to CD3 engagement: clues to a possible mechanism of anti-CD3 action?; J Immunol; 2013.
24574339	Keerthivasan S, Aghajani K, Dose M, et al.; ß-Catenin promotes colitis and colon cancer through imprinting of proinflammatory properties in T cells; Sci Transl Med; 2014.
Bioproject PRJNA117149	Ley T, Cai S; Expression data of Naive Treg, allogeneic tumor-activated Treg, and GVHD-activated Treg cells;2009 Aug 31;https://www.ncbi.nlm.nih.gov/bioproject/PRJNA117149 (online:2015 Dec 30)
Bioproject PRJNA143315	Chen X, Sharma A, Cyran A, et al.; Integrin alpha-6 (CD49f) defines a novel and distinct subset of CD4+ regulatory T cells with potent suppression activity.;2008 Jul 08;http://www.ncbi.nlm.nih.gov/bioproject/PRJNA143315 (online:2015 Dec 30)

Bibliography

[1] A. K. Abbas, A. H. Lichtman, and S. Pillai. *Cellular and molecular Immunology*. Elsevier Saunders, 8th edition edition, 2014.

[2] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398, 2010.

[3] H. Ahlfors, A. Limaye, L. L. Elo, S. Tuomela, M. Burute, K. V. Gottimukkala, D. Notani, O. Rasool, S. Galande, R. Lahesmaa, S. Anders, and W. Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.

[4] T. Äijö, S. M. Edelman, T. Lönnberg, A. Larjo, H. Kallionpää, S. Tuomela, E. Engström, R. Lahesmaa, and H. Lähdesmäki. An integrative computational systems biology approach identifies differentially regulated dynamic transcriptome signatures which drive the initiation of human t helper cell differentiation. *BMC genomics*, 13(1):1, 2012.

[5] M. F. Akay. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert systems with applications*, 36(2):3240–3247, 2009.

[6] T. Akimova, U. H. Beier, L. Wang, M. H. Levine, and W. W. Hancock. Helios expression is a marker of t cell activation and proliferation. *PLoS ONE*, 6(8):e24226, 2011.

[7] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Essential cell biology*. Garland Science, 2013.

[8] L. G. Almeida, N. J. Sakabe, M. C. C. Silva, A. S. Mundstein, T. Cohen, Y.-T. Chen, R. Chua, S. Gurung, S. Gnjatic, A. A. Jungbluth, et al. Ctdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic acids research*, 37(suppl 1):D816–D819, 2009.

[9] G. Altay and F. Emmert-Streib. Inferring the conservative causal core of gene regulatory networks. *BMC Systems Biology*, 4(1):1–13, 2010.

[10] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

[11] S. Andrews et al. Fastqc: A quality control tool for high throughput sequence data. *Reference Source*, 2010.

[12] M. Ante, E. Wingender, and M. Fuchs. Integration of gene expression data with prior knowledge for network analysis and validation. *BMC Res Notes*, 4:520, 2011.

[13] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, May 2000. ISSN 1061-4036. URL `http://dx.doi.org/10.1038/75556`.

[14] J. L. Badano and N. Katsanis. Beyond mendel: an evolving view of human genetic disease transmission. *Nat Rev Genet*, 3(10):779–789, Oct 2002.

[15] G. D. Bader, D. Betel, and C. W. V. Hogue. Bind: the biomolecular interaction network database. *Nucleic Acids Research*, 31(1):248–250, 2003.

[16] F. O. Bagger, D. Sasivarevic, S. H. Sohi, L. G. Laursen, S. Pundhir, C. K. Sønderby, O. Winther, N. Rapin, and B. T. Porse. Bloodspot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis. *Nucleic acids research*, 44(D1):D917–D924, 2016.

[17] S. Balaji, L. M. Iyer, M. M. Babu, and L. Aravind. Comparison of transcription regulatory interactions inferred from high-throughput methods: what do they reveal? *Trends in Genetics*, 24(7):319–323, 2008.

[18] P. Baldi and S. Brunak. *Bioinformatics: the machine learning approach*. MIT press, 2001.

[19] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo. How to infer gene networks from expression profiles. *Mol Syst Biol*, 3:–, Feb. 2007. URL `http://dx.doi.org/10.1038/msb4100120`.

[20] Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol*, 21(11):1337–1342, Nov 2003.

[21] A.-L. Barabási and E. Bonabeau. Scale-free networks. *Scientific American*, 288(5):50–59, 2003.

[22] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, et al. Ncbi geo: archive for functional genomics data sets—10 years on. *Nucleic acids research*, 39(suppl 1):D1005–D1010, 2011.

[23] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. Ncbi geo: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013.

[24] A. Ben-Hur and J. Weston. A user's guide to support vector machines. *Data mining techniques for the life sciences*, pages 223–239, 2010.

[25] M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C. M. Perou, and J. S. Marron. Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1):105–114, 2004.

[26] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Res*, 41(Database issue):D36–D42, Jan 2013.

[27] N. Berestovsky and L. Nakhleh. An evaluation of methods for inferring boolean networks from time-series data. *PLoS ONE*, 8(6):e66031, 06 2013.

[28] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[29] A. Bernard and A. J. Hartemink. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. *Pac Symp Biocomput*, pages 459–470, 2005.

[30] J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian statistics*, 7:733–742, 2003.

[31] D. P. Berrar, W. Dubitzky, M. Granzow, et al. *A practical approach to microarray data analysis*. Springer, 2003.

[32] T. Bo and I. Jonassen. New feature subset selection procedures for classification of expression profiles. *Genome biology*, 3(4):1–0017, 2002.

[33] H. Bolouri. Modeling genomic regulatory networks with big data. *Trends in Genetics*, 30(5):182 – 191, 2014.

[34] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[35] R. Bonecchi, G. Bianchi, P. P. Bordignon, D. D'Ambrosio, R. Lang, A. Borsatti, S. Sozzani, P. Allavena, P. A. Gray, A. Mantovani, and F. Sinigaglia. Differential expression of chemokine receptors and chemotactic responsiveness of type 1 t helper cells (th1s) and th2s. *J Exp Med*, 187 (1):129–34, 1998.

[36] R. Bonneau, D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. S. Baliga, and V. Thorsson. The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome biology*, 7(5):R36, 2006.

[37] M. Borenstein, L. V. Hedges, J. Higgins, and H. R. Rothstein. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2):97–111, 2010.

[38] S. G. Bottcher and C. Dethlefsen. *deal: Learning Bayesian Networks with Mixed Variables*, 2013. URL `http://CRAN.R-project.org/package=deal`. R package version 1.2-37.

[39] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, et al. Minimum information about a microarray experiment (miame)—toward standards for microarray data. *Nature genetics*, 29(4):365–371, 2001.

[40] K. Breuer, A. K. Foroushani, M. R. Laird, C. Chen, A. Sribnaia, R. Lo, G. L. Winsor, R. E. W. Hancock, F. S. L. Brinkman, and D. J. Lynn. Innatedb: systems biology of innate immunity and beyond–recent updates and continuing curation. *Nucleic Acids Res*, 41(Database issue):D1228–D1233, Jan 2013.

[41] C. C. Brinkman, D. Iwami, M. K. Hritzo, Y. Xiong, S. Ahmad, T. Simon, K. L. Hippen, B. R. Blazar, and J. S. Bromberg. Treg engage lymphotoxin beta receptor for afferent lymphatic transendothelial migration. *Nat Commun*, 7:12021, 2016.

[42] D. Bruder, M. Probst-Kepper, A. M. Westendorf, R. Geffers, S. Beissert, K. Loser, H. von Boehmer, J. Buer, and W. Hansen. Frontline: Neuropilin-1: a surface marker of regulatory t cells. *European journal of immunology*, 34(3):623–630, 2004.

[43] C. Brunner, A. Sindrilaru, I. Girkontaite, K. D. Fischer, C. Sunderkotter, and T. Wirth. Bob.1/obf.1 controls the balance of th1 and th2 immune responses. *EMBO J*, 26(13):3191–202, 2007.

[44] D. Busse, M. de la Rosa, K. Hobiger, K. Thurley, M. Flossdorf, A. Scheffold, and T. Höfer. Competing feedback loops shape il-2 signaling between helper and regulatory t lymphocytes in cellular microenvironments. *Proc Natl Acad Sci U S A*, 107(7):3058–3063, Feb 2010.

[45] A. J. Butte and I. S. Kohane. Unsupervised knowledge discovery in medical databases using relevance networks. page 711, 1999.

[46] A. J. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. 5:418–429, 2000.

[47] P. Cahan, F. Rovegno, D. Mooney, J. C. Newman, G. S. Laurent, and T. A. McCaffrey. Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. *Gene*, 401(1-2):12–18, Oct 2007.

[48] A. Campain and Y. H. Yang. Comparison study of microarray meta-analysis methods. *BMC Bioinformatics*, 11:408, 2010.

[49] D. J. Campbell and M. A. Koch. Phenotypical and functional specialization of foxp3+ regulatory t cells. *Nature Reviews Immunology*, 11(2):119–130, 2011.

[50] P. J. Campbell, S. Yachida, L. J. Mudie, P. J. Stephens, E. D. Pleasance, L. A. Stebbings, L. A. Morsberger, C. Latimer, S. McLaren, M.-L. Lin, et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, 467(7319):1109–1113, 2010.

[51] M. Caridade, L. Graca, and R. M. Ribeiro. Mechanisms underlying cd4+ treg immune regulation in the adult: from experiments to models. *Immune system modeling and analysis*, page 328, 2015.

[52] S. L. Carter, C. M. Brechbühler, M. Griffin, and A. T. Bond. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14):2242–2250, Sep 2004.

[53] L. E. Chai, S. K. Loh, S. T. Low, M. S. Mohamad, S. Deris, and Z. Zakaria. A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine*, 48(0):55 – 65, 2014.

[54] C. Chang, Z. Ding, Y. S. Hung, and P. C. W. Fung. Fast network component analysis (fastnca) for gene regulatory network reconstruction from microarray data. *Bioinformatics*, 24(11):1349–1358, Jun 2008.

[55] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[56] L.-C. Chang, H.-M. Lin, E. Sibille, and G. Tseng. Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinformatics*, 14(1):368, 2013.

[57] E. Chautard, L. Ballut, N. Thierry-Mieg, and S. Ricard-Blum. Matrixdb, a database focused on extracellular protein-protein and protein-carbohydrate interactions. *Bioinformatics*, 25(5):690–691, Mar 2009.

[58] N. V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1–6, 2004.

[59] W. Chen and J. E. Konkel. Development of thymic foxp3+ regulatory t cells: Tgf-β matters. *European journal of immunology*, 45(4):958–965, 2015.

[60] W. Chen, W. Jin, N. Hardegen, K.-j. Lei, L. Li, N. Marinos, G. McGrady, and S. M. Wahl. Conversion of peripheral cd4+ cd25- naive t cells to cd4+ cd25+ regulatory t cells by tgf-β induction of transcription factor foxp3. *The Journal of experimental medicine*, 198(12):1875–1886, 2003.

[61] Z. Chen, A. Laurence, and J. J. O'Shea. Signal transduction pathways and transcriptional regulation in the control of th17 differentiation. *Semin Immunol*, 19(6):400–408, Dec 2007.

[62] W.-C. Cheng, M.-L. Tsai, C.-W. Chang, C.-L. Huang, C.-R. Chen, W.-Y. Shu, Y.-S. Lee, T.-H. Wang, J.-H. Hong, C.-Y. Li, and I. C. Hsu. Microarray meta-analysis database (m2db): a uniformly pre-processed, quality controlled, and manually curated human clinical microarray database. *BMC Bioinformatics*, 11(1):1–9, 2010.

[63] H. Choi, R. Shen, A. M. Chinnaiyan, and D. Ghosh. A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC bioinformatics*, 8(1):364, 2007.

[64] J. K. Choi, U. Yu, S. Kim, and O. J. Yoo. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19(suppl 1):i84–i90, 2003.

[65] A. M. Chumakov, A. Silla, E. A. Williamson, and H. P. Koeffler. Modulation of dna binding properties of ccaat/enhancer binding protein epsilon by heterodimer formation and interactions with nfkappab pathway. *Blood*, 109(10):4209–19, 2007.

[66] M. Ciofani, A. Madar, C. Galan, M. Sellars, K. Mace, F. Pauli, A. Agarwal, W. Huang, C. N. Parkurst, M. Muratet, K. M. Newberry, S. Meadows, A. Greenfield, Y. Yang, P. Jain, F. K. Kirigin, C. Birchmeier, E. F. Wagner, K. M. Murphy, R. M. Myers, R. Bonneau, and D. R. Littman. A validated regulatory network for th17 cell specification. *Cell*, 151(2):289–303, 2012.

[67] L. Clarke, X. Zheng-Bradley, R. Smith, E. Kulesha, C. Xiao, I. Toneva, B. Vaughan, D. Preuss, R. Leinonen, M. Shumway, et al. The 1000 genomes project: data management and community access. *Nature methods*, 9(5):459–462, 2012.

[68] E. M. Conlon, J. J. Song, and A. Liu. Bayesian meta-analysis models for microarray data: a comparative study. *BMC Bioinformatics*, 8:80, 2007.

[69] E. P. Consortium et al. The encode (encyclopedia of dna elements) project. *Science*, 306(5696): 636–640, 2004.

[70] N. R. Coordinators. Database resources of the national center for biotechnology information. *Nucleic acids research*, 41(Database issue):D8, 2013.

[71] X. Z. L. Cope, E. Garrett, and G. Parmigiani. *MergeMaid: Merge Maid*, 2007. URL `http://astor.som.jhmi.edu/MergeMaid`. R package version 2.36.0.

[72] L. Corcoran, D. Emslie, T. Kratina, W. Shi, S. Hirsch, N. Taubenheim, and S. Chevrier. Oct2 and obf1 as facilitators of b:t cell collaboration during a humoral immune response. *Front Immunol*, 5:108, 2014.

[73] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[74] M. P. Cox, D. A. Peterson, and P. J. Biggs. Solexaqa: At-a-glance quality assessment of illumina second-generation sequencing data. *BMC Bioinformatics*, 11:485, 2010.

[75] S. Crotty. A brief history of t cell help to b cells. *Nature Reviews Immunology*, 15(3):185–189, 2015.

[76] d. L. M. Curotto and J. Lafaille. Natural and adaptive foxp3+ regulatory t cells: more of the same or a division of labor? *Immunity*, 30(5):626–635, 2009.

[77] L. F. da Rocha Junior, M. J. Rego, M. B. Cavalcanti, M. C. Pereira, M. G. Pitta, P. S. de Oliveira, S. M. Goncalves, A. L. Duarte, C. de Lima Mdo, R. Pitta Ida, and M. G. Pitta. Synthesis of a novel thiazolidinedione and evaluation of its modulatory effect on ifn- gamma , il-6, il-17a, and il-22 production in pbmcs from rheumatoid arthritis patients. *Biomed Res Int*, 2013:926060, 2013.

[78] D. D'Ambrosio, A. Iellem, R. Bonecchi, D. Mazzeo, S. Sozzani, A. Mantovani, and F. Sinigaglia. Selective up-regulation of chemokine receptors ccr4 and ccr8 upon activation of polarized human type 2 th cells. *J Immunol*, 161(10):5111–5, 1998.

[79] M. H. Daves, S. G. Hilsenbeck, C. C. Lau, and T.-K. Man. Meta-analysis of multiple microarray datasets reveals a common gene signature of metastasis in solid tumors. *BMC medical genomics*, 4(1):1, 2011.

[80] S. Davis and P. Meltzer. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 14:1846–1847, 2007.

[81] A. de la Fuente. From 'differential expression'to 'differential networking'–identification of dysfunctional regulatory networks in diseases. *Trends in genetics*, 26(7):326–333, 2010.

[82] R. de Matos Simoes and F. Emmert-Streib. Bagging statistical network inference from large-scale gene expression data. *PLoS ONE*, 7(3):e33624, 2012.

[83] R. de Matos Simoes and F. Emmert-Streib. *bc3net: Gene Regulatory Network Inference with Bc3net*, 2015. URL `https://CRAN.R-project.org/package=bc3net`. R package version 1.0.3.

[84] D. J. Depew and B. H. Weber. Darwinism evolving: Systems dynamics and the genealogy of natural selection. 1995.

[85] D. di Bernardo, M. J. Thompson, T. S. Gardner, S. E. Chobot, E. L. Eastwood, A. P. Wojtovich, S. J. Elliott, S. E. Schaus, and J. J. Collins. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat Biotechnol*, 23(3):377–383, Mar 2005.

[86] R. Díaz-Uriarte and S. A. De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):1, 2006.

[87] R. Dobrin, Q. K. Beg, A.-L. Barabási, and Z. N. Oltvai. Aggregation of topological motifs in the escherichia coli transcriptional regulatory network. *BMC Bioinformatics*, 5:10, Jan 2004.

[88] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. D. Moor, A. Brazma, and W. Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, Aug 2005.

[89] S. Durinck, P. T. Spellman, E. Birney, and W. Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nat Protoc*, 4(8):1184–91, 2009.

[90] P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, 2000.

[91] eBioscience Inc. Cytokine atlas, t helper 2 (th2) cells & cytokines. online, January 2016. URL `http://www.ebioscience.com/knowledge-center/antigen/cytokines/th2-cells.htm`.

[92] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868, Dec 1998.

[93] B. Elser, M. Lohoff, S. Kock, M. Giaisi, S. Kirchhoff, P. H. Krammer, and M. Li-Weber. Ifn-gamma represses il-4 expression via irf-1 and irf-2. *Immunity*, 17(6):703–12, 2002.

[94] E. Estrada and J. A. Rodríguez-Velázquez. Subgraph centrality in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 71(5 Pt 2):056103, May 2005.

[95] B. Everitt. *The Cambridge dictionary of statistics*. Cambridge Univ. Press, Cambridge [u.a.], 1998. ISBN 0521593468.

[96] I. Ezkurdia, D. Juan, J. M. Rodriguez, A. Frankish, M. Diekhans, J. Harrow, J. Vazquez, A. Valencia, and M. L. Tress. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum Mol Genet*, 23(22):5866–5878, Nov 2014.

[97] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5(1):e8, 01 2007.

[98] S. Falcon and R. Gentleman. Using gostats to test gene lists for go term association. *Bioinformatics*, 23(2):257–258, Jan 2007.

[99] Z. Fang, K. Hecklau, F. Gross, I. Bachmann, M. Venzke, M. Karl, J. Schuchhardt, A. Radbruch, H. Herzel, and R. Baumgrass. Transcription factor co-occupied regions in the murine genome constitute t-helper-cell subtype-specific enhancers. *European journal of immunology*, 45(11):3150–3157, 2015.

[100] J. D. Farrar, J. D. Smith, T. L. Murphy, S. Leung, G. R. Stark, and K. M. Murphy. Selective loss of type i interferon-induced stat4 activation caused by a minisatellite insertion in mouse stat2. *Nat Immunol*, 1(1):65–9, 2000.

[101] M. Feuerer, Y. Shen, D. R. Littman, C. Benoist, and D. Mathis. How punctual ablation of regulatory t cells unleashes an autoimmune lesion within the pancreatic islets. *Immunity*, 31(4): 654–664, 2009.

[102] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, and L. J. Jensen. String v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*, 41(Database issue):D808–D815, Jan 2013.

[103] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[104] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29 (2-3):131–163, 1997.

[105] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

[106] S. Gallagher, S. E. Winston, S. A. Fuller, and J. G. R. Hurrell. Immunoblotting and immunodetection. *Curr Protoc Mol Biol*, Chapter 10:Unit 10.8, Jul 2008.

[107] C. A. Gallo, R. L. Cecchini, J. A. Carballido, S. Micheletto, and I. Ponzoni. Discretization of gene expression data revised. *Briefings in Bioinformatics*, 2015.

[108] T. S. Gardner and J. J. Faith. Reverse-engineering transcription control networks. *Physics of Life Reviews*, 2(1):65 – 88, 2005.

[109] T. S. Gardner, D. di Bernardo, D. Lorenz, and J. J. Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301(5629):102–105, Jul 2003.

[110] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[111] J. Geginat, M. Paroni, S. Maglie, J. S. Alfen, I. Kastirr, P. Gruarin, M. De Simone, M. Pagani, and S. Abrignani. Plasticity of human cd4 t cell subsets. *CD4+ T cell differentiation in infection: amendments to the Th1/Th2 axiom*, page 67, 2015.

[112] F. Geier, J. Timmer, and C. Fleck. Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC Syst Biol*, 1:11, 2007.

[113] M. Ghanbari, J. Lasserre, and M. Vingron. Reconstruction of gene networks using prior knowledge. *BMC Syst Biol*, 9:84, 2015.

[114] D. Ghosh and H. Choi. *metaArray: Integration of Microarray Data for Meta-analysis*, 2014. URL `https://www.bioconductor.org/packages//2.7/bioc/html/metaArray.html`. R package version 1.40.0.

[115] F. M. Giorgi, C. Del Fabbro, and F. Licausi. Comparative study of rna-seq-and microarray-derived coexpression networks in arabidopsis thaliana. *Bioinformatics*, 29(6):717–724, 2013.

[116] J. Gollub, C. A. Ball, G. Binkley, J. Demeter, D. B. Finkelstein, J. M. Hebert, T. Hernandez-Boussard, H. Jin, M. Kaloper, J. C. Matese, et al. The stanford microarray database: data access and quality assessment tools. *Nucleic Acids Research*, 31(1):94–96, 2003.

[117] J. A. Goodacre and G. Dick. *Immunopathogenetic mechanisms of arthritis*. Springer Science & Business Media, 2012.

[118] A. Graessel, S. M. Hauck, C. von Toerne, E. Kloppmann, T. Goldberg, H. Koppensteiner, M. Schindler, B. Knapp, L. Krause, K. Dietz, et al. A combined omics approach to generate the surface atlas of human naive cd4+ t cells during early t-cell receptor activation. *Molecular & Cellular Proteomics*, 14(8):2085–2102, 2015.

[119] A. Greenfield, C. Hafemeister, and R. Bonneau. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, 29(8):1060–1067, 2013.

[120] M. Grimaldi, R. Visintainer, and G. Jurman. Regnann: reverse engineering gene networks using artificial neural networks. *PloS one*, 6(12):e28646, 2011.

[121] G. R. Grimes, S. Moodie, J. S. Beattie, M. Craigon, P. Dickinson, T. Forster, A. D. Livingston, M. Mewissen, K. A. Robertson, A. J. Ross, et al. Gpx-macrophage expression atlas: A database for expression profiles of macrophages challenged with a variety of pro-inflammatory, anti-inflammatory, benign and pathogen insults. *BMC genomics*, 6(1):1, 2005.

[122] D. GuhaThakurta and G. D. Stormo. Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17(7):608–621, Jul 2001.

[123] Z. Guo, B. Tzvetkova, J. M. Bassik, T. Bodziak, B. M. Wojnar, W. Qiao, M. A. Obaida, S. B. Nelson, B. H. Hu, and P. Yu. Rnaseqmetadb: a database and web server for navigating metadata of publicly available mouse rna-seq datasets. *Bioinformatics*, 31(24):4038–4040, 2015.

[124] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[125] H. Hache, H. Lehrach, and R. Herwig. Reverse engineering of gene regulatory networks: a comparative study. *EURASIP Journal on Bioinformatics and Systems Biology*, 2009(1):1–12, 2009.

[126] J. B. Hagen. The origins of bioinformatics. *Nat Rev Genet*, 1(3):231–236, Dec 2000.

[127] S. Haider and A. K. Zaidi. Transforming timed influence nets into time sliced bayesian networks. Technical report, DTIC Document, 2004.

[128] M. Hall. A decision tree-based attribute weighting filter for naive bayes. *Knowledge-Based Systems*, 20(2):120–126, 2007.

[129] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11.1, 2009.

[130] M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, Hamilton, New Zealand., 1998.

[131] S. Han, J. Nam, Y. Li, S. Kim, S. H. Cho, Y. S. Cho, S. Y. Choi, J. Choi, K. Han, Y. Kim, M. Na, H. Kim, Y. C. Bae, S. Y. Choi, and E. Kim. Regulation of dendritic spines, spatial memory, and embryonic development by the tanc family of psd-95-interacting proteins. *J Neurosci*, 30(45): 15102–12, 2010.

[132] B. Harr and C. Schlötterer. Comparison of algorithms for the analysis of affymetrix microarray data as evaluated by co-expression of genes in known operons. *Nucleic Acids Research*, 34(2): e8–e8, 2006.

[133] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[134] W. M. Hartmann. Dimension reduction vs. variable selection. In *Applied Parallel Computing. State of the Art in Scientific Computing*, pages 931–938. Springer, 2004.

[135] F. He, J. Buer, A.-P. Zeng, and R. Balling. Dynamic cumulative activity of transcription factors as a mechanism of quantitative gene regulation. *Genome Biol*, 8(9):R181, 2007.

[136] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28, 1998.

[137] M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke. Gene regulatory network inference: Data integration in dynamic models–a review. *Biosystems*, 96(1):86 – 103, 2009.

[138] D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

[139] T. Heinemeyer, E. Wingender, I. Reuter, H. Hermjakob, A. E. Kel, O. Kel, E. V. Ignatieva, E. A. Ananko, O. A. Podkolodnaya, F. Kolpakov, et al. Databases on transcriptional regulation: Transfac, trrd and compel. *Nucleic acids research*, 26(1):362–367, 1998.

[140] N. D. Heintzman, G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, C. W. Ching, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, 2009.

[141] S. Hempel, A. Koseska, Z. Nikoloski, and J. Kurths. Unraveling gene regulatory networks from time-resolved gene expression data - a measures comparison study. *BMC Bioinformatics*, 12:292, 2011.

[142] A. Hinks, S. Eyre, X. Ke, A. Barton, P. Martin, E. Flynn, J. Packham, J. Worthington, S. Childhood Arthritis Prospective, U. Consortium, B. S. Group, and W. Thomson. Association of the aff3 gene and il2/il21 gene region with juvenile idiopathic arthritis. *Genes Immun*, 11(2):194–8, 2010.

[143] Z. M. Hira and D. F. Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Adv Bioinformatics*, 2015:198363, 2015.

[144] P. Hogeweg. The roots of bioinformatics in theoretical biology. *PLoS Comput Biol*, 7(3):e1002021, Mar 2011.

[145] K. Honda, H. Yanai, H. Negishi, M. Asagiri, M. Sato, T. Mizutani, N. Shimada, Y. Ohba, A. Takaoka, N. Yoshida, and T. Taniguchi. Irf-7 is the master regulator of type-i interferon-dependent immune responses. *Nature*, 434(7034):772–7, 2005.

[146] F. Hong and R. Breitling. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24(3):374–382, Feb 2008.

[147] F. Hong, R. Breitling, C. W. McEntee, B. S. Wittner, J. L. Nemhauser, and J. Chory. Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–2827, Nov 2006.

[148] A. Honkela, J. Peltonen, H. Topa, I. Charapitsa, F. Matarese, K. Grote, H. G. Stunnenberg, G. Reid, N. D. Lawrence, and M. Rattray. Genome-wide modeling of transcription kinetics reveals patterns of rna production delays. *Proc Natl Acad Sci U S A*, 112(42):13115–13120, Oct 2015.

[149] S. Hori, T. Nomura, and S. Sakaguchi. Control of regulatory t cell development by the transcription factor foxp3. *Science*, 299(5609):1057–1061, 2003.

[150] V. Hower, S. N. Evans, and L. Pachter. Shape-based peak identification for chip-seq. *BMC Bioinformatics*, 12:15, 2011.

[151] C.-S. Hsieh, Y. Zheng, Y. Liang, J. D. Fontenot, and A. Y. Rudensky. An intersection between the self-reactive regulatory and nonregulatory t cell receptor repertoires. *Nature immunology*, 7 (4):401–410, 2006.

[152] C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

[153] G. Hu, Q. Tang, S. Sharma, F. Yu, T. M. Escobar, S. A. Muljo, J. Zhu, and K. Zhao. Expression and regulation of intergenic long noncoding rnas during t cell development and differentiation. *Nat Immunol*, 14(11):1190–8, 2013.

[154] E. Huala, A. W. Dickerman, M. Garcia-Hernandez, D. Weems, L. Reiser, F. LaFond, D. Hanley, D. Kiphart, M. Zhuang, W. Huang, et al. The arabidopsis information resource (tair): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic acids research*, 29(1):102–105, 2001.

[155] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009.

[156] H. Huang, Y. Ma, W. Dawicki, X. Zhang, and J. R. Gordon. Comparison of induced versus natural regulatory t cells of the same tcr specificity for induction of tolerance to an environmental antigen. *The Journal of Immunology*, 191(3):1136–1143, 2013.

[157] N. J. Hudson, A. Reverter, P. L. Greenwood, B. Guo, L. M. Cafe, and B. P. Dalrymple. Longitudinal muscle gene expression patterns associated with differential intramuscular fat in cattle. *Animal*, 9(4):650–9, 2015.

[158] J. Huehn, K. Siegmund, J. C. Lehmann, C. Siewert, U. Haubold, M. Feuerer, G. F. Debes, J. Lauber, O. Frey, G. K. Przybylski, et al. Developmental stage, phenotype, and migration distinguish naive-and effector/memory-like cd4+ regulatory t cells. *The Journal of experimental medicine*, 199(3):303–313, 2004.

[159] S. Hunter, M. Corbett, H. Denise, M. Fraser, A. Gonzalez-Beltran, C. Hunter, P. Jones, R. Leinonen, C. McAnulla, E. Maguire, et al. Ebi metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucleic acids research*, 42(D1):D600–D606, 2014.

[160] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9), 2010.

[161] T. Ideker, J. Dutkowski, and L. Hood. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell*, 144(6):860–863, 2011.

[162] K. Imamura, S. Maeda, I. Kawamura, K. Matsuyama, N. Shinohara, Y. Yahiro, S. Nagano, T. Setoguchi, M. Yokouchi, Y. Ishidou, and S. Komiya. Human immunodeficiency virus type 1 enhancer-binding protein 3 is essential for the expression of asparagine-linked glycosylation 2 in the regulation of osteoblast and chondrocyte differentiation. *J Biol Chem*, 289(14):9865–79, 2014.

[163] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, T. P. Speed, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[164] O. Irsoy, O. T. Yildiz, and E. Alpaydin. Design and analysis of classifier learning experiments in bioinformatics: survey and case studies. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 9(6):1663–1675, 2012.

[165] R. Jabeen, R. Goswami, O. Awe, A. Kulkarni, E. T. Nguyen, A. Attenasio, D. Walsh, M. R. Olson, M. H. Kim, R. S. Tepper, J. Sun, C. H. Kim, E. J. Taparowsky, B. Zhou, and M. H. Kaplan. Th9 cell development requires a batf-regulated transcriptional network. *J Clin Invest*, 123(11):4641–53, 2013.

[166] A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(2):153–158, 1997.

[167] C. A. Janeway, P. Travers, M. Walport, M. J. Shlomchik, et al. *Immunobiology: the immune system in health and disease*, volume 2. Garland New York, 2001.

[168] M. Jargosch, S. Kröger, E. Gralinska, U. Klotz, Z. Fang, W. Chen, U. Leser, J. Selbig, D. Groth, and R. Baumgrass. Data integration for identification of important transcription factors of stat6-mediated cell fate decisions. *Genet Mol Res*, 15(2), 2016.

[169] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 2001.

[170] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: a survey. *Knowledge and Data Engineering, IEEE Transactions on*, 16(11):1370–1386, 2004.

[171] W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[172] S. Z. Josefowicz, R. E. Niec, H. Y. Kim, P. Treuting, T. Chinen, Y. Zheng, D. T. Umetsu, and A. Y. Rudensky. Extrathymically generated regulatory t cells control mucosal th2 inflammation. *Nature*, 482(7385):395–399, 2012.

[173] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl 1):D428–D432, 2005.

[174] A. R. Joyce and B. O. Palsson. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol*, 7(3):198–210, Mar. 2006. ISSN 1471-0072. URL `http://dx.doi.org/10.1038/nrm1857`.

[175] A. Kamburov, U. Stelzl, H. Lehrach, and R. Herwig. The consensuspathdb interaction database: 2013 update. *Nucleic Acids Res*, 41(Database issue):D793–D800, Jan 2013.

[176] M. Kanehisa. The kegg database. *silico simulation of biological processes*, 247:91–103, 2002.

[177] C. Kanz, P. Aldebert, N. Althorpe, W. Baker, A. Baldwin, K. Bates, P. Browne, A. van den Broek, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, J. Gamble, F. G. Diez, N. Harte, T. Kulikova, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M. McHale, F. Nardone, V. Silventoinen, S. Sobhany, P. Stoehr, M. A. Tuli, K. Tzouvara, R. Vaughan, D. Wu, W. Zhu, and R. Apweiler. The embl nucleotide sequence database. *Nucleic Acids Res*, 33(Database issue): D29–D33, Jan 2005.

[178] A. Karnowski, S. Chevrier, G. T. Belz, A. Mount, D. Emslie, K. D'Costa, D. M. Tarlinton, A. Kallies, and L. M. Corcoran. B and t cells collaborate in antiviral responses via il-6, il-21, and transcriptional activator and coactivator, oct2 and obf-1. *J Exp Med*, 209(11):2049–64, 2012.

[179] M. Kashiwada, S. L. Cassel, J. D. Colgan, and P. B. Rothman. Nfil3/e4bp4 controls type 2 t helper cell cytokine expression. *EMBO J*, 30(10):2071–82, 2011.

[180] A. Kauffmann, R. Gentleman, and W. Huber. arrayqualitymetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–416, Feb 2009.

[181] S. Keerthikumar, R. Raju, K. Kandasamy, A. Hijikata, S. Ramabadran, L. Balakrishnan, M. Ahmed, S. Rani, L. D. N. Selvan, D. S. Somanathan, et al. Rapid: resource of asian primary immunodeficiency diseases. *Nucleic acids research*, 37(suppl 1):D863–D867, 2009.

[182] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob. The intact molecular interaction database in 2012. *Nucleic Acids Res*, 40(Database issue):D841–D846, Jan 2012.

[183] B. L. Kidder, G. Hu, and K. Zhao. Chip-seq: technical considerations for obtaining high-quality data. *Nature immunology*, 12(10):918–922, 2011.

[184] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14(4):R36, 2013.

[185] K. Kira and L. A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *AAAI*, volume 2, pages 129–134, 1992.

[186] H. T. Kissick, M. G. Sanda, L. K. Dunn, K. L. Pellegrini, S. T. On, J. K. Noel, and M. S. Arredouani. Androgens alter t-cell immunity by inhibiting t-helper 1 differentiation. *Proc Natl Acad Sci U S A*, 111(27):9887–92, 2014.

[187] J. Kittler. *Feature selection and extraction*. 1986.

[188] U. Kjærulff. dhugin: A computational system for dynamic time-sliced bayesian networks. *International journal of forecasting*, 11(1):89–111, 1995.

[189] Y. Kodama, J. Mashima, E. Kaminuma, T. Gojobori, O. Ogasawara, T. Takagi, K. Okubo, and Y. Nakamura. The dna data bank of japan launches a new resource, the ddbj omics archive of functional genomics experiments. *Nucleic Acids Res*, 40(Database issue):D38–D42, Jan 2012.

[190] T. Koga, C. M. Hedrich, M. Mizui, N. Yoshida, K. Otomo, L. A. Lieberman, T. Rauen, J. C. Crispin, and G. C. Tsokos. Camk4-dependent activation of akt/mtor and crem-alpha underlies autoimmunity-associated th17 imbalance. *J Clin Invest*, 124(5):2234–45, 2014.

[191] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1): 273–324, 1997.

[192] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[193] R. Konig, S. Stertz, Y. Zhou, A. Inoue, H. H. Hoffmann, S. Bhattacharyya, J. G. Alamares, D. M. Tscherne, M. B. Ortigoza, Y. Liang, Q. Gao, S. E. Andrews, S. Bandyopadhyay, P. De Jesus, B. P. Tu, L. Pache, C. Shih, A. Orth, G. Bonamy, L. Miraglia, T. Ideker, A. Garcia-Sastre, J. A. Young, P. Palese, M. L. Shaw, and S. K. Chanda. Human host factors required for influenza virus replication. *Nature*, 463(7282):813–7, 2010.

[194] S. Konishi, T. Ando, and S. Imoto. Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, 91(1):27–43, 2004.

[195] T. Korn, E. Bettelli, M. Oukka, and V. K. Kuchroo. Il-17 and th17 cells. *Annual review of immunology*, 27:485–517, 2009.

[196] S. Kothari, J. H. Phan, T. H. Stokes, A. O. Osunkoya, A. N. Young, and M. D. Wang. Removing batch effects from histopathological images for enhanced cancer diagnosis. *Biomedical and Health Informatics, IEEE Journal of*, 18(3):765–772, 2014.

[197] M. Kris Wetterstrand. Dna sequencing costs data from the nhgri genome sequencing program (gsp). online, January 2015. URL `http://www.genome.gov/sequencingcosts/`.

[198] S. Kroeger, M. Venzke, M. Karl, R. Baumgrass, and U. Leser. Transcriptome-based discrimination of treg cell subtypes: culture conditions matter. <in preparation>, November 2016.

[199] S. J. Kwon, J. Crespo-Barreto, W. Zhang, T. Wang, D. S. Kim, A. Krensky, and C. Clayberger. Klf13 cooperates with c-maf to regulate il-4 expression in cd4+ t cells. *J Immunol*, 192(12):5703–9, 2014.

[200] E. S. Lander. Array of hope. *Nature genetics*, 21:3–4, 1999.

[201] P. Langfelder and S. Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1, 2008.

[202] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9 (4):357–359, 2012.

[203] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112, 2006.

[204] A. Laurence, S. Amarnath, J. Mariotti, Y. C. Kim, J. Foley, M. Eckhaus, J. J. O'Shea, and D. H. Fowler. Stat3 transcription factor promotes instability of ntreg cells and limits generation of itreg cells during acute murine graft-versus-host disease. *Immunity*, 37(2):209–22, 2012.

[205] Y. Lee, A. C. Scheck, T. F. Cloughesy, A. Lai, J. Dong, H. K. Farooqi, L. M. Liau, S. Horvath, P. S. Mischel, and S. F. Nelson. Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age. *BMC medical genomics*, 1(1):52, 2008.

[206] Y. K. Lee, R. Mukasa, R. D. Hatton, and C. T. Weaver. Developmental plasticity of th17 and treg cells. *Curr Opin Immunol*, 21(3):274–280, Jun 2009.

[207] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28 (6):882–883, 2012.

[208] R. Lehmann, L. Childs, P. Thomas, M. Abreu, L. Fuhr, H. Herzel, U. Leser, and A. Relógio. Assembly of a comprehensive regulatory network for the mammalian circadian clock: a bioinformatics approach. *PloS one*, 10(5):e0126283, 2015.

[209] R. Leinonen, R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tárraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, et al. The european nucleotide archive. *Nucleic acids research*, page gkq967, 2010.

[210] R. Leinonen, H. Sugawara, and M. Shumway. The sequence read archive. *Nucleic acids research*, page gkq1019, 2010.

[211] B. Lemaitre, E. Nicolas, L. Michaut, J. M. Reichhart, and J. A. Hoffmann. The dorsoventral regulatory gene cassette spätzle/Toll/cactus controls the potent antifungal response in Drosophila adults. *Cell*, 86(6):973–983, Sep 1996.

[212] J. Z. Levin, M. Yassour, X. Adiconis, C. Nusbaum, D. A. Thompson, N. Friedman, A. Gnirke, and A. Regev. Comprehensive comparative analysis of strand-specific rna sequencing methods. *Nature methods*, 7(9):709–715, 2010.

[213] I. Levner. Feature selection and nearest centroid classification for protein mass spectrometry. *BMC bioinformatics*, 6(1):1, 2005.

[214] G. Li, H. Peng, J. Zhang, and L. Zhu. Robust rank correlation based screening. *The Annals of Statistics*, pages 1846–1877, 2012.

[215] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009.

[216] H. Li, D. Zhu, and M. Cook. A statistical framework for consolidating ßiblingpprobe sets for affymetrix genechip data. *BMC Genomics*, 9:188, 2008.

[217] J. Li. *Statistical Issues in Meta-analysis for Identifying Signature Genes in the Integration of Multiple Genomic Studies*. ProQuest, 2008.

[218] Y. Li and L. Chen. Big biological data: challenges and opportunities. *Genomics Proteomics Bioinformatics*, 12(5):187–189, Oct 2014.

[219] Y. Li and S. A. Jackson. Gene network reconstruction by integration of prior biological knowledge. *G3 (Bethesda)*, 5(6):1075–1079, Jun 2015.

[220] Z. Li, P. Li, A. Krishnan, and J. Liu. Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic bayesian network analysis. *Bioinformatics*, 27(19):2686–2691, 2011.

[221] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput*, pages 18–29, 1998.

[222] J. C. Liao, R. Boscolo, Y.-L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci U S A*, 100(26):15522–15527, Dec 2003.

[223] D. D. Licatalosi and R. B. Darnell. Rna processing and its regulation: global insights into biological networks. *Nature Reviews Genetics*, 11(1):75–87, 2010.

[224] S.-W. Lin, Z.-J. Lee, S.-C. Chen, and T.-Y. Tseng. Parameter determination of support vector machine and feature selection using simulated annealing approach. *Applied soft computing*, 8(4): 1505–1512, 2008.

[225] R. Lister, B. D. Gregory, and J. R. Ecker. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Current opinion in plant biology*, 12(2):107–118, 2009.

[226] C.-C. Liu, J. Hu, M. Kalakrishnan, H. Huang, and X. Zhou. Integrative disease classification based on cross-platform microarray data. *BMC Bioinformatics*, 10(1):1–8, 2009.

[227] J. Y. Liu, F. Li, L. P. Wang, X. F. Chen, D. Wang, L. Cao, Y. Ping, S. Zhao, B. Li, S. H. Thorne, B. Zhang, P. Kalinski, and Y. Zhang. Ctl- vs treg lymphocyte-attracting chemokines, ccl4 and ccl20, are strong reciprocal predictive markers for survival of patients with oesophageal squamous cell carcinoma. *Br J Cancer*, 113(5):747–755, Sep 2015.

[228] Q. Liu, A. H. Sung, Z. Chen, J. Liu, X. Huang, and Y. Deng. Feature selection and classification of maqc-ii breast cancer and multiple myeloma microarray gene expression data. *PloS ONE*, 4 (12):e8250, 2009.

[229] Q. Liu, A. H. Sung, M. Qiao, Z. Chen, J. Y. Yang, M. Q. Yang, X. Huang, and Y. Deng. Comparison of feature selection and classification for maldi-ms data. *BMC genomics*, 10(1):1, 2009.

[230] Q. Liu, A. Boudot, J. Ni, T. Hennessey, S. L. Beauparlant, H. N. Rajabi, C. Zahnow, and M. E. Ewen. Cyclin d1 and c/ebpbeta lap1 operate in a common pathway to promote mammary epithelial cell differentiation. *Mol Cell Biol*, 34(16):3168–79, 2014.

[231] X. Liu, R. I. Nurieva, and C. Dong. Transcriptional regulation of follicular t-helper (tfh) cells. *Immunol Rev*, 252(1):139–45, 2013.

[232] Y. Liu, A. An, and X. Huang. Boosting prediction accuracy on imbalanced datasets with svm ensembles. pages 107–118, 2006.

[233] Z. Liu, H. Fan, and S. Jiang. Cd4(+) t-cell subsets in transplantation. *Immunol Rev*, 252(1): 183–191, Mar 2013.

[234] Z.-P. Liu. Reverse engineering of genome-wide gene regulatory networks from gene expression data. *Curr Genomics*, 16(1):3–22, Feb 2015.

[235] M. Lohoff and T. W. Mak. Roles of interferon-regulatory factors in t-helper-cell differentiation. *Nat Rev Immunol*, 5(2):125–35, 2005.

[236] T. A. Long, S. M. Brady, and P. N. Benfey. Systems approaches to identifying gene regulatory networks in plants. *Annu Rev Cell Dev Biol*, 24:81–103, 2008.

[237] J. Lu, J. C. Lee, M. L. Salit, and M. C. Cam. Transcript-based redefinition of grouped oligonucleotide probe sets using aceview: high-resolution annotation for microarrays. *BMC bioinformatics*, 8(1):108, 2007.

[238] R. Lund, H. Ahlfors, E. Kainonen, A. M. Lahesmaa, C. Dixon, and R. Lahesmaa. Identification of genes involved in the initiation of human th1 or th2 cell commitment. *Eur J Immunol*, 35(11): 3307–19, 2005.

[239] L. Lusa, R. Gentleman, and M. Ruschhaupt. *GeneMeta: MetaAnalysis for High Throughput Experiments*, 2013. URL `https://www.bioconductor.org/packages/release/bioc/html/GeneMeta.html`. R package version 1.34.0.

[240] J. Ma, R. Wang, X. Fang, Y. Ding, and Z. Sun. Critical role of tcf-1 in repression of the il-17 gene. *PLoS ONE*, 6(9):e24768, 2011.

[241] S. Ma, J. Sung, A. T. Magis, Y. Wang, D. Geman, and N. D. Price. Measuring the effect of inter-study variability on estimating prediction error. *PLoS ONE*, 9(10):e110840, 2014.

[242] K. L. MacQuarrie, A. P. Fong, R. H. Morse, and S. J. Tapscott. Genome-wide transcription factor binding: beyond direct target regulation. *Trends in Genetics*, 27(4):141 – 148, 2011.

[243] A. Madar, A. Greenfield, H. Ostrer, E. Vanden-Eijnden, and R. Bonneau. The inferelator 2.0: a scalable framework for reconstruction of dynamic regulatory network models. *Conf Proc IEEE Eng Med Biol Soc*, 2009:5448–5451, 2009.

[244] K. J. Maloy and F. Powrie. Fueling regulation: Il-2 keeps cd4+ treg cells fit. *Nat Immunol*, 6(11): 1071–1072, Nov 2005.

[245] K. J. Mantione, R. M. Kream, H. Kuzelova, R. Ptacek, J. Raboch, J. M. Samuel, and G. B. Stefano. Comparing bioinformatic gene expression profiling methods: microarray and rna-seq. *Medical science monitor basic research*, 20:138, 2014.

[246] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proc Natl Acad Sci U S A*, 107 (14):6286–6291, Apr 2010.

[247] D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, G. Stolovitzky, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.

[248] S. Marco. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, 35(3), 2010.

[249] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7, 2006.

[250] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18(9): 1509–1517, Sep 2008.

[251] M. N. McCall, B. M. Bolstad, and R. A. Irizarry. Frozen robust multiarray analysis (frma). *Biostatistics*, 11(2):242–253, 2010.

[252] A. Meixner, F. Karreth, L. Kenner, and E. F. Wagner. Jund regulates lymphocyte proliferation and t helper cell cytokine expression. *The EMBO journal*, 23(6):1325–1335, 2004.

[253] P. Meyer, D. Marbach, S. Roy, and M. Kellis. Information-theoretic inference of gene networks using backward elimination. In *BIOCOMP*, pages 700–705, 2010.

[254] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics and systems biology*, 2007(1):1–9, 2007.

[255] H. Mi, B. Lazareva-Ulitsky, R. Loo, A. Kejariwal, J. Vandergriff, S. Rabkin, N. Guo, A. Muruganujan, O. Doremieux, M. J. Campbell, et al. The panther database of protein families, subfamilies, functions and pathways. *Nucleic acids research*, 33(suppl 1):D284–D288, 2005.

[256] M. Miyara, Y. Ito, and S. Sakaguchi. Treg-cell therapies for autoimmune rheumatic diseases. *Nature Reviews Rheumatology*, 2014.

[257] M. T. Morgan. Rnaseq analysis. online, 2014. URL `http://bioconductor.fhcrc.org/help/course-materials/2014/SeattleFeb2014/`.

[258] N. Morshed, M. Chetty, and N. X. Vinh. Fusgp: bayesian co-learning of gene regulatory networks and protein interaction networks. pages 369–377, 2012.

[259] K. Murphy, S. Mian, et al. Modelling gene expression data using dynamic bayesian networks. 1999. Technical report, Computer Science Division, University of California, Berkeley, CA.

[260] K. Murphy et al. The bayes net toolbox for matlab. *Computing science and statistics*, 33(2): 1024–1034, 2001.

[261] C. Müssel, M. Hopfensitz, D. Zhou, H. A. Kestler, A. Biere, T. D. Hanson, and M. H. A. Kestler. *BoolNet*, 2015. URL `https://cran.r-project.org/web/packages/BoolNet/index.html`. R package version 2.1.3.

[262] N. C. B. I Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 44(D1):D7–D19, Jan 2016.

[263] R. Nagarajan, M. Scutari, and S. Lebre. *Bayesian Networks in R with Applications in Systems Biology*. Springer, New York, 2013.

[264] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, and M. J. Bamshad. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*, 42(1):30–35, Jan 2010.

[265] H. Nishikawa and S. Sakaguchi. Regulatory t cells in cancer immunotherapy. *Current opinion in immunology*, 27:1–7, 2014.

[266] S.-L. T. Normand. Tutorial in biostatistics meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in medicine*, 18(3):321–359, 1999.

[267] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni, and H. Hermjakob. The mintact project–intact as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res*, 42(Database issue):D358–D363, Jan 2014.

[268] J. J. O'Shea and W. E. Paul. Mechanisms underlying lineage commitment and plasticity of helper cd4+ t cells. *Science*, 327(5969):1098–1102, Feb 2010.

[269] J. Owen, J. Punt, and S. Stranford. *Kuby Immunology*. W.H. Freeman, 7th revised international edition edition, 2013.

[270] S. K. Pabbisetty, W. Rabacal, D. Maseda, D. Cendron, P. L. Collins, K. L. Hoek, V. V. Parekh, T. M. Aune, and E. Sebzda. Klf2 is a rate-limiting transcription factor that can be targeted to enhance regulatory t-cell production. *Proc Natl Acad Sci U S A*, 111(26):9579–84, 2014.

[271] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stümpflen, H.-W. Mewes, et al. The mips mammalian protein–protein interaction database. *Bioinformatics*, 21(6):832–834, 2005.

[272] A. Palazon, A. W. Goldrath, V. Nizet, and R. S. Johnson. Hif transcription factors, inflammation, and immunity. *Immunity*, 41(4):518–28, 2014.

[273] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 0-934613-73-7.

[274] C. A. Penfold and D. L. Wild. How to infer gene networks from expression profiles, revisited. *Interface Focus*, 1(6):857–870, 2011.

[275] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[276] S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. K. B. Gandhi, K. N. Chandrika, N. Deshpande, S. Suresh, B. P. Rashmi, K. Shanker, N. Padma, V. Niranjan, H. C. Harsha, N. Talreja, B. M. Vrushabendra, M. A. Ramya, A. J. Yatish, M. Joy, H. N. Shivashankar, M. P. Kavitha, M. Menezes, D. R. Choudhury, N. Ghosh, R. Saravana, S. Chandran, S. Mohan, C. K. Jonnalagadda, C. K. Prasad, C. Kumar-Sinha, K. S. Deshpande, and A. Pandey. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, 32 (Database issue):D497–D501, Jan 2004.

[277] L. E. Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[278] C. Petzold, N. Steinbronn, M. Gereke, R. H. Strasser, T. Sparwasser, D. Bruder, R. Geffers, S. Schallenberg, and K. Kretschmer. Fluorochrome-based definition of naturally occurring foxp3+ regulatory t cells of intra-and extrathymic origin. *European journal of immunology*, 44(12):3632–3645, 2014.

[279] J. C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Advances in kernel methods - support vector learning, 1998.

[280] B. R. Powell, N. R. Buist, and P. Stenzel. An x-linked syndrome of diarrhea, polyendocrinopathy, and fatal infection in infancy. *The Journal of pediatrics*, 100(5):731–737, 1982.

[281] P. Praveen and H. Fröhlich. Boosting probabilistic graphical model inference by incorporating prior knowledge from multiple sources. *PLoS ONE*, 8(6):e67410, 2013.

[282] M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers. *BMC genomics*, 13(1):1, 2012.

[283] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[284] N. Raghavachari, J. Barb, Y. Yang, P. Liu, K. Woodhouse, D. Levy, C. J. O'Donnell, P. J. Munson, and G. J. Kato. A systematic comparison and evaluation of high density exon arrays and rna-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. *BMC Med Genomics*, 5:28, 2012.

[285] A. Ramasamy, A. Mondry, C. C. Holmes, and D. G. Altman. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med*, 5(9):e184, Sep 2008.

[286] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26):15149–15154, 2001.

[287] A. Rambaut, O. G. Pybus, M. I. Nelson, C. Viboud, J. K. Taubenberger, and E. C. Holmes. The genomic and epidemiological dynamics of human influenza a virus. *Nature*, 453(7195):615–619, 2008.

[288] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, Aug 2002.

[289] J. Rengarajan, B. Tang, and L. H. Glimcher. Nfatc2 and nfatc3 regulate t(h)2 differentiation and modulate tcr-responsiveness of naive t(h)cells. *Nat Immunol*, 3(1):48–54, 2002.

[290] D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan. Meta-analysis of microarrays interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer research*, 62(15):4427–4433, 2002.

[291] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pander, and A. M. Chinnaiyan. Oncomine: a cancer microarray database and integrated data-mining platform. *Neoplasia*, 6(1):1–6, 2004.

[292] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A*, 101(25):9309–9314, Jun 2004.

[293] C. Ribeiro de Almeida, H. Heath, S. Krpic, G. M. Dingjan, J. P. van Hamburg, I. Bergen, S. van de Nobelen, F. Sleutels, F. Grosveld, N. Galjart, and R. W. Hendriks. Critical role for the transcription regulator ccctc-binding factor in the control of th2 cytokine expression. *J Immunol*, 182(2):999–1010, 2009.

[294] M. W. Richardson, J. Jadlowsky, C. A. Didigu, R. W. Doms, and J. L. Riley. Kruppel-like factor 2 modulates ccr5 expression and susceptibility to hiv-1 infection. *J Immunol*, 189(8):3815–21, 2012.

[295] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896–902, 2014.

[296] B. Ristevski. A survey of models for inference of gene regulatory networks. *Nonlinear Anal Model Control*, 18(4):444–65, 2013.

[297] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[298] E. M. Ross, D. Bourges, T. V. Hogan, P. A. Gleeson, and I. R. Driel. Helios defines t cells being driven to tolerance in the periphery and thymus. *European journal of immunology*, 2014.

[299] J. M. Rozenberg, P. Bhattacharya, R. Chatterjee, K. Glass, and C. Vinson. Combinatorial recruitment of creb, c/ebpbeta and c-jun determines activation of promoters upon keratinocyte differentiation. *PLoS ONE*, 8(11):e78179, 2013.

[300] B. E. Russ, J. E. Prier, S. Rao, and S. J. Turner. T cell immunity as a tool for studying epigenetic regulation of cellular differentiation. *Epigenetic Modifications and Viral Infections*, page 25, 2015.

[301] R. L. Rutishauser, G. A. Martins, S. Kalachikov, A. Chandele, I. A. Parish, E. Meffre, J. Jacob, K. Calame, and S. M. Kaech. Transcriptional repressor blimp-1 promotes cd8+ t cell terminal differentiation and represses the acquisition of central memory t cell properties. *Immunity*, 31(2):296–308, 2009.

[302] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[303] A. Saito, S. Kanemoto, Y. Zhang, R. Asada, K. Hino, and K. Imaizumi. Chondrocyte proliferation regulated by secreted luminal domain of er stress transducer bbf2h7/creb3l2. *Mol Cell*, 53(1):127–39, 2014.

[304] S. Sakaguchi, N. Sakaguchi, M. Asano, M. Itoh, and M. Toda. Immunologic self-tolerance maintained by activated t cells expressing il-2 receptor alpha-chains (cd25). breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. *The Journal of Immunology*, 155 (3):1151–1164, 1995.

[305] G. Sales and C. Romualdi. *parmigene: Parallel Mutual Information estimation for Gene Network reconstruction.*, 2012. URL `http://CRAN.R-project.org/package=parmigene`. R package version 1.0.2.

[306] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Res*, 32(Database issue):D449–D451, Jan 2004.

[307] A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(suppl 1): D91–D94, 2004.

[308] F. Sanger. Nucleotide sequence of bacteriophage (d x174 dna. 1977.

[309] J. Saric, L. J. Jensen, R. Ouzounova, I. Rojas, and P. Bork. Extraction of regulatory gene/protein networks from medline. *Bioinformatics*, 22(6):645–650, Mar 2006.

[310] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467, 1995.

[311] F. L. Schmidt and J. E. Hunter. *Methods of meta-analysis: Correcting error and bias in research findings*. Sage publications, 2014.

[312] E. G. Schmitt and C. B. Williams. Generation and function of induced regulatory t cells. *Frontiers in immunology*, 4, 2013.

[313] D. E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, 2008.

[314] B. U. Schraml, K. Hildner, W. Ise, W. L. Lee, W. A. Smith, B. Solomon, G. Sahota, J. Sim, R. Mukasa, S. Cemerski, R. D. Hatton, G. D. Stormo, C. T. Weaver, J. H. Russell, T. L. Murphy, and K. M. Murphy. The ap-1 transcription factor batf controls t(h)17 differentiation. *Nature*, 460 (7253):405–9, 2009.

[315] M. H. Schulz, W. E. Devanny, A. Gitter, S. Zhong, J. Ernst, and Z. Bar-Joseph. Drem 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC Syst Biol*, 6:104, 2012.

[316] M. Scutari. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*, 2009.

[317] A. A. Shabalin, H. Tjelmeland, C. Fan, C. M. Perou, and A. B. Nobel. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, 24(9):1154–1160, 2008.

[318] A. Shameli, J. Yamanouchi, S. Tsai, Y. Yang, X. Clemente-Casares, A. Moore, P. Serra, and P. Santamaria. Il-2 promotes the function of memory-like autoregulatory cd8+ t cells but suppresses their development via foxp3+ treg cells. *Eur J Immunol*, 43(2):394–403, Feb 2013.

[319] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[320] T. Shay and J. Kang. Immunological genome project and systems immunology. *Trends in immunology*, 34(12):602–609, 2013.

[321] J. Shendure and H. Ji. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

[322] S. Sherry. Ncbi sra toolkit technology for next generation sequence data. In *Plant and Animal Genome XX Conference (January 14-18, 2012)*. Plant and Animal Genome, 2012.

[323] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.

[324] E. M. Shevach and A. M. Thornton. ttregs, ptregs, and itregs: similarities and differences. *Immunological reviews*, 259(1):88–102, 2014.

[325] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2):261–274, 2002.

[326] A. H. Sims, G. J. Smethurst, Y. Hey, M. J. Okoniewski, S. D. Pepper, A. Howell, C. J. Miller, and R. B. Clarke. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets–improving meta-analysis and prediction of prognosis. *BMC medical genomics*, 1(1):1, 2008.

[327] B. D. Singer, L. S. King, and F. R. D'Alessio. Regulatory t cells as immunotherapy. *Frontiers in immunology*, 5, 2014.

[328] J. Sławek and T. Arodź. Adanet: inferring gene regulatory networks using ensemble classifiers. *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 434–441, 2012.

[329] G. K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

[330] G. K. Smyth and T. Speed. Normalization of cdna microarray data. *Methods*, 31(4):265–273, 2003.

[331] F. Spitz and E. E. Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, 2012.

[332] P. Stafford. *Methods in microarray normalization*. CRC Press, 2008.

[333] M. A. Stalteri and A. P. Harrison. Interpretation of multiple probe sets mapping to the same gene in affymetrix genechips. *BMC bioinformatics*, 8(1):13, 2007.

[334] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Res*, 34(Database issue):D535–D539, Jan 2006.

[335] E. Steele. *Combining heterogeneous sources of data for the reverse-engineering of gene regulatory networks*. PhD thesis, School of Information Systems, Computing and Mathematics, Brunel University, January 2010.

[336] E. Steele, A. Tucker, P. A. C. 't Hoen, and M. J. Schuemie. Literature-based priors for gene regulatory networks. *Bioinformatics*, 25(14):1768–1774, Jul 2009.

[337] L. D. Stein et al. The case for cloud computing in genome informatics. *Genome Biol*, 11(5):207, 2010.

[338] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[339] G. Stoesser, W. Baker, A. van den Broek, M. Garcia-Pastor, C. Kanz, T. Kulikova, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, F. Nardone, P. Stoehr, M. A. Tuli, K. Tzouvara, and R. Vaughan. The embl nucleotide sequence database: major new developments. *Nucleic Acids Res*, 31(1):17–22, Jan 2003.

[340] G. L. Stritesky, R. Muthukrishnan, S. Sehra, R. Goswami, D. Pham, J. Travers, E. T. Nguyen, D. E. Levy, and M. H. Kaplan. The transcription factor stat3 is required for t helper 2 cell development. *Immunity*, 34(1):39–49, 2011.

[341] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *science*, 302(5643):249–255, 2003.

[342] A. Sturn, J. Quackenbush, and Z. Trajanoski. Genesis: cluster analysis of microarray data. *Bioinformatics*, 18(1):207–208, 2002.

[343] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, Oct 2005.

[344] S. L. Swain, K. K. McKinstry, and T. M. Strutt. Expanding roles for $CD4^+$ T cells in immunity to viruses. *Nat Rev Immunol*, 12(2):136–148, Feb 2012.

[345] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. von Mering. String v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*, 43 (Database issue):D447–D452, Jan 2015.

[346] A. Sîrbu, H. J. Ruskin, and M. Crane. Cross-platform microarray data normalisation for regulatory network inference. *PLoS ONE*, 5(11):e13822, 11 2010.

[347] J. Taminau, S. Meganck, C. Lazar, D. Steenhoff, A. Coletta, C. Molter, R. Duque, V. de Schaetzen, D. Y. W. Solís, H. Bersini, et al. Unlocking the potential of publicly available microarray data using insilicodb and insilicomerging r/bioconductor packages. *BMC bioinformatics*, 13(1):335, 2012.

[348] R. J. Tan, L. J. Gibbons, C. Potter, K. L. Hyrich, A. W. Morgan, A. G. Wilson, J. D. Isaacs, Braggss, and A. Barton. Investigation of rheumatoid arthritis susceptibility genes identifies association of aff3 and cd226 variants with response to anti-tumour necrosis factor treatment. *Ann Rheum Dis*, 69(6):1029–35, 2010.

[349] S. Tanaka, A. Suto, T. Iwamoto, D. Kashiwakuma, S. Kagami, K. Suzuki, H. Takatori, T. Tamachi, K. Hirose, A. Onodera, J. Suzuki, O. Ohara, M. Yamashita, T. Nakayama, and H. Nakajima. Sox5 and c-maf cooperatively induce th17 cell differentiation via rorgammat induction as downstream targets of stat3. *J Exp Med*, 211(9):1857–74, 2014.

[350] Q. Tang, J. A. Bluestone, and S.-M. Kang. Cd4+ foxp3+ regulatory t cell therapy in transplantation. *Journal of molecular cell biology*, 4(1):11–21, 2012.

[351] Y. Tang, Y.-Q. Zhang, and Z. Huang. Development of two-stage svm-rfe gene selection strategy for microarray expression data analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 4(3):365–381, 2007.

[352] C. M. Tato and J. J. O'Shea. Immunology: what does it mean to be just 17? *Nature*, 441(7090): 166–168, 2006.

[353] R. C. Taylor, A. Shah, C. Treatman, and M. Blevins. Sebini: software environment for biological network inference. *Bioinformatics*, 22(21):2706–2708, 2006.

[354] D. Tenenbaum and B. Team. *BiocInstaller: Install/Update Bioconductor and CRAN Packages*, 2016. URL `http://bioconductor.org/packages/3.2/BiocViews.html`. R package version 1.18.5.

[355] A. Theil, S. Tuve, U. Oelschlägel, A. Maiwald, D. Döhler, D. Oßmann, A. Zenkel, C. Wilhelm, J. M. Middeke, N. Shayegi, et al. Adoptive transfer of allogeneic regulatory t cells into patients with chronic graft-versus-host disease. *Cytotherapy*, 2015.

[356] P. Thomas. *Robust relationship extraction in the biomedical domain*. PhD thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät, 2015.

[357] P. Thomas, M. Neves, I. Solt, D. Tikk, and U. Leser. Relation Extraction for Drug-Drug Interactions using Ensemble Learning. In *Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction 2011*, pages 11–18, 2011.

[358] P. Thomas, I. Solt, R. Klinger, and U. Leser. Learning Protein Protein Interaction Extraction using Distant Supervision. In *Proceedings of Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing*, pages 34–41, 2011.

[359] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562–578, 2012.

[360] I. Tsamardinos, C. F. Aliferis, A. R. Statnikov, and E. Statnikov. Algorithms for large scale markov blanket discovery. In *FLAIRS conference*, volume 2, 2003.

[361] G. C. Tseng, D. Ghosh, and E. Feingold. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res*, Jan 2012.

[362] H. R. Ueda, W. Chen, A. Adachi, H. Wakamatsu, S. Hayashi, T. Takasugi, M. Nagano, K.-i. Nakahama, Y. Suzuki, S. Sugano, et al. A transcription factor response element for gene expression during circadian night. *Nature*, 418(6897):534–539, 2002.

[363] U.S. National Library of Medicine. Geo documentation. online, 2015. URL `http://www.ncbi.nlm.nih.gov/geo/info/`.

[364] T. Ushijima, K. Okazaki, H. Tsushima, and Y. Iwamoto. Ccaat/enhancer-binding protein beta regulates the repression of type ii collagen expression during the differentiation from proliferative to hypertrophic chondrocytes. *J Biol Chem*, 289(5):2852–63, 2014.

[365] G. Vahedi, H. Takahashi, S. Nakayamada, H. W. Sun, V. Sartorelli, Y. Kanno, and J. J. O'Shea. Stats shape the active enhancer landscape of t cell populations. *Cell*, 151(5):981–93, 2012.

[366] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263, 2009.

[367] K. V. Voelkerding, S. A. Dames, and J. D. Durtschi. Next-generation sequencing: from basic research to diagnostics. *Clin Chem*, 55(4):641–658, Apr 2009.

[368] H. von Boehmer and C. Daniel. Therapeutic opportunities for manipulating treg cells in autoimmunity and cancer. *Nature reviews Drug discovery*, 12(1):51–63, 2013.

[369] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork. String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 33(Database issue):D433–D437, Jan 2005.

[370] K. Wang and H. Nishida. Regulator: a database of metazoan transcription factors and maternal factors for developmental studies. *BMC bioinformatics*, 16(1):1, 2015.

[371] X. Wang, X. Liu, S. Matwin, and N. Japkowicz. Applying instance-weighted support vector machines to class imbalanced datasets. pages 112–118, 2014.

[372] X. V. Wang, N. Blades, J. Ding, R. Sultana, and G. Parmigiani. Estimation of sequencing error rates in short reads. *BMC Bioinformatics*, 13:185, 2012.

[373] Y. K. Wang, D. G. Hurley, S. Schnell, C. G. Print, and E. J. Crampin. Integration of steady-state and temporal gene expression data for the inference of gene regulatory networks. *PLoS ONE*, 8 (8):e72103, 2013.

[374] Z. Wang, W. Xu, F. A. San Lucas, and Y. Liu. Incorporating prior knowledge into gene network study. *Bioinformatics*, 29(20):2633–2640, Oct 2013.

[375] P. Warnat, R. Eils, and B. Brors. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 6:265, 2005.

[376] S. Wasserman and K. Faust. *Social network analysis : methods and applications*. Cambridge University Press, Cambridge; New York, 1994.

[377] W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287, 2004.

[378] G. Wei, L. Wei, J. Zhu, C. Zang, J. Hu-Li, Z. Yao, K. Cui, Y. Kanno, T.-Y. Roh, W. T. Watford, D. E. Schones, W. Peng, H.-w. Sun, W. E. Paul, J. J. O'Shea, and K. Zhao. Global mapping of h3k4me3 and h3k27me3 reveals specificity and plasticity in lineage fate determination of differentiating cd4+ t cells. *Immunity*, 30(1):155–167, Jan. 2009. ISSN 1074-7613.

[379] L. Wei, G. Vahedi, H. W. Sun, W. T. Watford, H. Takatori, H. L. Ramos, H. Takahashi, J. Liang, G. Gutierrez-Cruz, C. Zang, W. Peng, J. J. O'Shea, and Y. Kanno. Discrete roles of stat4 and stat6 transcription factors in tuning epigenetic modifications and transcription during t helper cell differentiation. *Immunity*, 32(6):840–51, 2010.

[380] J. M. Weiss, A. M. Bilate, M. Gobert, Y. Ding, M. A. C. de Lafaille, C. N. Parkhurst, H. Xiong, J. Dolpady, A. B. Frey, M. G. Ruocco, et al. Neuropilin 1 is expressed on thymus-derived natural regulatory t cells, but not mucosa-generated induced foxp3+ t reg cells. *The Journal of experimental medicine*, 209(10):1723–1742, 2012.

[381] A. V. Werhli and D. Husmeier. Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat Appl Genet Mol Biol*, 6:Article15, 2007.

[382] P. Wong and E. G. Pamer. CD8 T cell responses to infectious pathogens. *Annu Rev Immunol*, 21: 29–70, 2003.

[383] J. Wu and J. B. Lingrel. Kruppel-like factor 2, a novel immediate-early transcriptional factor, regulates il-2 expression in t lymphocyte activation. *J Immunol*, 175(5):3060–6, 2005.

[384] M. Wuelling, F. J. Kaiser, L. A. Buelens, D. Braunholz, R. A. Shivdasani, R. Depping, and A. Vortkamp. Trps1, a regulator of chondrocyte proliferation and differentiation, interacts with the activator form of gli3. *Dev Biol*, 328(1):40–53, 2009.

[385] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research*, 30(1):303–305, 2002.

[386] M. Xiong, X. Fang, and J. Zhao. Biomarker identification by feature wrappers. *Genome Research*, 11(11):1878–1887, 2001.

[387] L. Xu, A. C. Tan, R. L. Winslow, and D. Geman. Merging microarray data from separate breast cancer studies provides a robust prognostic test. *Bmc Bioinformatics*, 9(1):125, 2008.

[388] M. Yadav, C. Louvet, D. Davini, J. M. Gardner, M. Martinez-Llordella, S. Bailey-Bucktrout, B. A. Anthony, F. M. Sverdrup, R. Head, D. J. Kuster, et al. Neuropilin-1 distinguishes natural and inducible regulatory t cells among regulatory t cell subsets in vivo. *The Journal of experimental medicine*, 209(10):1713–1722, 2012.

[389] M. Yadav, S. Stephan, and J. A. Bluestone. Peripherally induced tregs–role in immune homeostasis and autoimmunity. *Frontiers in immunology*, 4, 2013.

[390] H. Yamane and W. E. Paul. Early signaling events that underlie fate decisions of naive cd4(+) t cells toward distinct t-helper cell subsets. *Immunol Rev*, 252(1):12–23, 2013.

[391] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420, 1997.

[392] Y. H. Yang, Y. Xiao, and M. R. Segal. Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics*, 21(7):1084–1093, 2005.

[393] K. Yeung and R. Bumgarner. Correction: Multiclass classification of microarray data with repeated measurements: application to cancer. *Genome biology*, 6(13):405, 2005.

[394] H. Yi, C. Guo, X. Yu, D. Zuo, and X. Y. Wang. Mouse cd11b+gr-1+ myeloid cells can promote th17 cell differentiation and experimental autoimmune encephalomyelitis. *J Immunol*, 189(9): 4295–304, 2012.

[395] N. Yosef, A. K. Shalek, J. T. Gaublomme, H. Jin, Y. Lee, A. Awasthi, C. Wu, K. Karwacz, S. Xiao, M. Jorgolli, D. Gennert, R. Satija, A. Shakya, D. Y. Lu, J. J. Trombetta, M. R. Pillai, P. J. Ratcliffe, M. L. Coleman, M. Bix, D. Tantin, H. Park, V. K. Kuchroo, and A. Regev. Dynamic regulatory network controlling th17 cell differentiation. *Nature*, 496(7446):461–8, 2013.

[396] W. C. Young, A. E. Raftery, and K. Y. Yeung. Fast bayesian inference for gene regulatory networks using scanbma. *BMC Syst Biol*, 8:47, 2014.

[397] M. Yousef, N. Najami, L. Abedallah, and W. Khalifa. Computational approaches for biomarker discovery. *Journal of Intelligent Learning Systems and Applications*, 6(4):153, 2014.

[398] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis. Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20 (18):3594–3603, Dec 2004.

[399] Q. Yu, A. Sharma, S. Y. Oh, H. G. Moon, M. Z. Hossain, T. M. Salay, K. E. Leeds, H. Du, B. Wu, M. L. Waterman, Z. Zhu, and J. M. Sen. T cell factor 1 initiates the t helper type 2 fate by inducing the transcription factor gata-3 and repressing interferon-gamma. *Nat Immunol*, 10(9): 992–9, 2009.

[400] X. Yu, D. Rollins, K. A. Ruhn, J. J. Stubblefield, C. B. Green, M. Kashiwada, P. B. Rothman, J. S. Takahashi, and L. V. Hooper. Th17 cell differentiation is regulated by the circadian clock. *Science*, 342(6159):727–30, 2013.

[401] B. Zacher, K. Abnaof, S. Gade, E. Younesi, A. Tresch, and H. Froehlich. *birta: Bayesian Inference of Regulation of Transcriptional Activity*, 2013. R package version 1.4.0.

[402] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*, 4:Article17, 2005.

[403] R. Zhang, K. Chen, L. Peng, and H. Xiong. Regulation of t helper cell differentiation by interferon regulatory factor family members. *Immunol Res*, 54(1-3):169–76, 2012.

[404] S. Zhao, W.-P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu. Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PLoS ONE*, 9(1):e78644, 2014.

[405] L. Zhou, M. M. W. Chong, and D. R. Littman. Plasticity of cd4+ t cell lineage differentiation. *Immunity*, 30(5):646–655, May 2009.

[406] J. Zhu and W. E. Paul. Peripheral cd4+ t-cell differentiation regulated by networks of cytokines and transcription factors. *Immunol Rev*, 238(1):247–62, 2010.

[407] J. Zhu, B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner, and E. E. Schadt. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet*, 40(7):854–861, Jul 2008.

[408] Y. Zhu, S. Davis, R. Stephens, P. S. Meltzer, and Y. Chen. Geometadb: powerful alternative search engine for the gene expression omnibus. *Bioinformatics*, 24(23):2798–2800, Dec 2008.

[409] E. Zintzaras and J. P. Ioannidis. Meta-analysis for ranked discovery datasets: theoretical framework and empirical demonstration for microarrays. *Computational biology and chemistry*, 32(1): 39–47, 2008.

[410] P. Zoppoli, S. Morganella, and M. Ceccarelli. Timedelay-aracne: Reverse engineering of gene networks from time-course data by an information theoretic approach. *Bmc Bioinformatics*, 11 (1):1, 2010.

[411] S. Žitnik, M. Žitnik, B. Zupan, and M. Bajec. Sieve-based relation extraction of gene regulatory networks from biological literature. *BMC Bioinformatics*, 16 Suppl 16:S1, 2015.

List of Figures

2.1. **Development of naïve T (Th0) cells into T helper cell lineages/ subtypes.** Process of T cell differentiation is mainly driven by presence and absence of lineage specific cytokines (written on the arcs) and so called master TFs (written in the cells). Cytokines at the right side are those, expressed by the activated cells. 10

2.2. **Putative mechanisms used by Treg cells.** Treg cells have multipe putative targets as described by Caridade et al. [51]. These include (1) targeting dendritic cells (DCs) – leading to weak or abrogated signals to naïve/effector T cells; (2) Metabolic disruption; (3) Competition – for critical cytokines, such as IL-2, or direct disruption of effector cell engagement with APCs; (4) Cytolysis – direct cytotoxic effect and consequent apoptosis of effector T cells or APCs; (5) Production of inhibitory cytokines – including IL-10, IL-35, and TGF-ß. Figure adapted from [51]. . 12

2.3. **Gene expression of eukaryotic organ is initiated in the cell nucleus.** DNA is transcribed into RNA. Next RNA is spliced and transported from cell nucleus to the cytoplasm and finally translated to a protein coding amino acid sequence. 13

2.4. **TF-binding in promoter region initiates gene transcription.** Binding of only one TF may cause a regulatory event, while binding of two can cause a different regulatory events [122]. The figure illustrates the phenomenon that gene regulation is cause by cooperatively acting of TFs, so called co-binding. 14

2.5. **Schema of a microarray experiment from sample purification to data analysis.** The figure illustrates processing steps and their sequential order during a microarray experiments. (Necessary sample preparation steps before purification are not shown here.) 15

2.6. **Microarray expression analysis workflow.** Various workflow variants similar to the presented (implementing different algorithms) exist. 16

2.7. **The decrease of sequencing costs per human genome over time.** The orange line additionally shows the cost per raw Megabases of DNA sequences. Data taken from the NHGRI Genome Sequencing Program (GSP) [197]. 17

2.8. **Illustration of the processing workflow for ChIP-seq experiments.** ChIP-seq procedure consists of i) the isolation of the target samples (here immune cells), ii) fragmentation of the isolated DNA, followed by iii) the chromatin immunoprecipitation with the subsequent RNA purification, iv) adapter ligation and end repair of the isolated sequences, v) library preparation and vi) clustering the targeted RNAs following by vii) the actual sequencing and subsequent analysis of detected reads. The figure is adapted from Kidder et al. [183]. 18

2.9. **Chromatin Immunoprecipitation sequencing(ChIP-seq) peak calling analysis workflow.** Analysis workflow consist of six steps followed by an additional annotation step (extraction of associated genes). While step 1-4 are similar to RNA-seq analysis, peak calling and peak annotation are specific to ChIP-seq analysis. 19

2.10. **RNA sequencing(RNA-seq) expression analysis workflow.**The steps 1-4 are similar to the ChIP-seq analysis workflow (Figure 2.8), while step 5-7 are specific to RNA-seq analysis. The general outline of the workflow is based on the work of Trapnell et al. [359] 20

3.1. **GEO Gene Omnibus platform growth over the last 15 years.** Plot illustrates the orders of magnitudes of platform, studies and samples submitted to the GEO database each year since inception. Until January 2016 about 1,669 million samples from 64,816 series (studies) perform on 15,358 were submitted to the GEO database. (Numbers were take from the website (`https://www.ncbi.nlm.nih.gov/geo/summary/?type=history`, online on 2016/01/10) 24

3.2. **Schema illustrates the steps of meta-analysis starting from requesting online repositories up to assembling the meta expression matrix.** Sections 3.4.1 - 3.4.6 describe the pipeline in detail. 32

3.3. **Fractions of platform types in the final meta expression set.** 154 arrays (29 iTreg and 125 nTreg) from 36 GEO series were selected. This set contained six different Affymetrix GeneChip®technologies. Numbers below the platform type indicate the corresponding amount of integrated microarrays. 34

3.4. **Different ProbeSet Ids link to the same Ensembl Mouse Gene Identifier** This problem is solved by assigning the expression values of the ProbeSet with the highest fold change between measured experimental states,(e.g. stimulation, cell type, treatment or time point) to the gene. . 36

3.5. **Hierarchical clustering of assembled meta expression matrix.** Most of the microarrays from the same experiment (GSE) cluster together, which illustrate the lab-/batch effect, as their expression profile is more similar among studies than towards other experiments. 37

3.6. **Upper boxplots show expression distribution of non-adjusted expression values after meta experiment assembly.** Varying expression values are clearly visible for both cell types (nTreg - red, iTreg - blue) and for all included studies. **Lower boxplots** show expression distribution after applying quantile re-normalization. Mean expression as well as minimal and maximal expression are adjusted to common levels, to enable subsequent expression analysis. 38

3.7. **QQplots illustrate difference before and after re-normalization. First row of figures** shows qqplots of non-adjusted data, where **second row of figures** shows re-normalized data. By comparing upper and lower figures it becomes clear that expression values became more homogeneous, which illustrates the reduction of the inter-study effect (the shift onto the diagonal), at the same time the profile of expression differences between single genes are widely kept (compressions and stretches in line profiles). 200 randomly selected genes were used to plot four comparisons of expression profiles from two independent microarrays (origin from different studies) each time. 40

3.8. **Genes differentially expressed between iTreg and nTreg cells.** Heatmap shows the mean gene expression extracted from the assembled and re-normalized meta expression set. Color key indicates expression values. Upper band of the heatmap represents expression in iTreg cells, where lower band represents expression in nTreg cells. 40

4.1. **Support Vector Machines (SVM) aim to find the "best" hyperplane (line a) that separates two classes.** Blue and yellow dots represent entities of two classes, that are separated by a line (a), that maximized decision boundaries (arrow-line d). Two dimension can be separated by a line, like illustrated here, while for higher-dimensional problems a hyperplane separates the classes. A small number of support vectors define the decision function. The support vectors line c_1 and c_2 determine the decision boundaries. Line b is one "not optimal" example of a separating hyperplane. 51

4.2. **Data preparation and feature selection pipeline.** Green arrows and braces represent data flows. We performed two data split steps to ensure that all samples from identical studies were not shared between data sets. Feature selection depicts the strongest attributes; here the genes that describe Treg cell subtype class best. GSE is accession number prefix of GEO series, that identifies a specific study. 52

4.3. **Frequency of simultaneously selected features (genes) by both methods on different subsets.** We extracted those genes as candidates that were selected as features in more than 50% (here: 18 of 36) of the subset analyses. This cut-off led to 41 candidate genes. 54

4.4. **First phase of evaluation: Classification of experiments based on potential marker genes.** The complete meta expression set was used for leave-one-out cross validation, but samples from the same study are not shared between training and test sets. 56

4.5. **Weighted performance metrics of SVM-based SMO classification.** Leave-one-out cross validation was performed. Features were restricted to **a)** 6 manually selected genes out of **b)** 41 genes as found by feature selection. For each test set the weighted mean was used to balance metrics derived for both Treg subtype classes. In contrast to Table 4.3, metrics were calculated for each test set classification separately to illustrate variances between the sets. 58

4.6. **a) Relative mRNA expression of candidate genes was analyzed by qRT-PCR.** mRNA was extracted from cultured iTreg and *ex vivo* nTreg cells of Foxp3-IRES-GFP mice and normalized to RPS-18 expression (n=6). The results show high consistency for predicted markers between meta-analysis and validation experiments; validation and original experiments were performed under similar conditions (non-cultivated nTreg cells). **b) Comparison of fold changes for candidates derived from meta-analysis and qRT-PCR experiments.** Rough-hatched bars show log_2 fold changes calculated from meta-analysis. Fine-hatched bars depict fold changes calculated from qRT-PCR between cultured iTreg and *ex vivo* nTreg cell. 59

4.7. **Immunoblot analysis of EMP1 and GABRR2.** Analysis revealed stability of differences in protein expression between cell subtypes for both candidates. Expressions of both proteins where subtracted by background signal and normalized to LaminB expression. 60

4.8. **Gene expression log_2 fold changes including cultured *ex vivo* nTreg cells.** Non-hatched bars show fold changes calculated from Bioinformatics meta-analysis. Rough-hatched bars depict fold changes calculated from qRT-PCR between iTreg and not cultured nTreg cell. Fine-hatched bars depict fold changes from qRT-PCR between iTreg and five days cultured nTreg cells. 61

4.9. **Number of microarray experiments per study.** The majority (72%) of integrated studies/ study parts are rather small, containing less than five microarrays. 62

5.1. **Directed GRN with three nodes v_1, v_2, v_3 and two edges e_1, e_2.** Node v_1 is the source of $e_1$and target of e_2, v_2 is targeted by a regulatory event initiated by v_1. v_1 might represent a TF that is target of another TF, here v_3. Regulatory strength can be represented as edge weights assigned to e_1 and e_2. 65

5.2. **A Bayesian network representing conditional probabilities among variables.** The graph illustrates a simple BN, where the $v_1 ... v_5$ represent genes and the causal relationship amongst them. In this example the expression of v_3 and v_4 depends on v_1 and v_2 respectively, where the expression of v_5 depends on the common expression of v_3 and v_4. Despite no direct relation between v_1 and v_5, it is clear, that v_5 depends on the expression of v_1 and this relation is mediated by v_3. 69

5.3. **The left network G_{DBN} represents the Dynamic Bayesian Network (DBN) over two time points of a time-series.** The transitions $v_1 \rightarrow v_3 \rightarrow v_5 \rightarrow v_1$ represent a circle transition over two time points. Right side BN illustrate the DBN on the left $G_{DBN} = G_{BN1} \cup G_{BN2}$, while split into two networks, one for each time-point. The mentioned circle is distributed over both subnetworks (G_{BN1}, G_{BN2}), as the only parent of $v_1 \in G_{BN2}$ is $v_5 \in G_{BN1}$ 71

5.4. **Tool testing strategy.** The six layers (1-6) illustrate the main steps of the pipeline in processing order. 84

5.5. **The line plots illustrate the tool specific characteristics of the performance metrics (y-axis) for increasing gene set sizes (x-axis).** From upper left to lower right the figures show F-measure, sensitivity, accuracy and specificity. Labels on the x-axis depict the gene set size. The measures were calculated based on the "combined" evaluation corpus. The lines for Banjo are plotted before the lines of Genie3 and thus fully or almost invisible. 88

5.6. **Boxplots of F-measure and accuracy aggregated for each tool over all gene sets to illustrate variance.** Comparison shows that no tool is able to outperform all others in terms of F-measure and accuracy. Even more we observed that higher number of correctly inferred interactions come along with a higher number of false positives. 89

6.1. **Integration strategy to create an interaction network around a master regulator (master TF).** This pipeline models the integration of four different data sources and is divided into three parts, i) data selection and preparation ii) analysis of each data set using specific workflows. iii) A mapping schema using a common id system is required to ensure compatibility between the data sets. Entities and links between the data sets are integrated into a single network based on unambiguous identifier only. 95

6.2. **Workflow for the construction of the STAT6 network.** The used filter criteria and data sources for each step are shown in the first row and highlighted in different colors (green for RNA-seq, purple for microarray data, orange for ChIP-seq). The filter criteria and/or the number of genes are listed for each integration step and gene group in the second row. Figure taken from corresponding publication [168]. 99

6.3. **The Venn-diagram shows the three filtered gene sets forming the basic network.** The intersection (100 genes, highlighted bold) of STAT6-regulated genes (green) with TFs, cytokines and cytokine receptors (purple) are dissected into direct and indirect target genes of STAT6 by integrating the STAT6 ChIP-seq data (orange). Figure adapted from [168]. 101
6.4. **Gene regulatory network of STAT6-regulated genes in Th2 cells.** The network is limited to TFs (square node), cytokines or cytokine receptors (oval node) and shows indirect (48 genes) and direct targets (52 genes) of STAT6. The network layout is based on four rings around STAT6 in graded gray tones. Rings from inside to outside contain genes according to their expression behavior: genes are only early (1), early and late (2), only late (3) and not (4) differentially expressed in Th2 cells compared to naïve $CD4^+$ T cells. The edge color indicates the effect of *Stat6* on gene expression changes, while the node color indicates the expression changes compared to naïve $CD4^+$ T cells. Figure adapted from [168]. 103
6.5. **Expression preferences of Th cell subtype-specific and STAT6-regulated genes.** Gene expression levels of 100 STAT6-regulated genes were compared between the Th cell subtypes Th1, Th2, Th17, and iTreg. Figure adapted from [168]. 104

7.1. **Number of bases uploaded to the Sequence Read Archive (SRA) [209].** The SRA together with the European Nucleotide Archive (ENA) [210] and the DNA Database of Japan [189] are the biggest open repository for sequence data. All three are partners in the International Nucleotide Sequence Database Collaboration (INSDC), that agreed on sharing all experiments. 110

List of Tables

3.1. **Listing of criteria for inclusion or rejection of experiments while manual curation by immunologist.** 33

3.2. **Table shows the size of the collected data set for each step of the pipeline.** The column "matched GEO-query" contains the numbers of elements as retrieved from GEO, while "selected" contains those, included in the meta expression data set after filtering and quality control. 34

3.3. **Overview of integrated gene expression arrays.** All microarrays are in situ oligonucleotide arrays manufactured by Affymetrix. 35

3.4. **Table illustrates expression value adjustment via re-normalization to improve comparability between data from different studies.** Table shows the means of all variances obtained for all genes in the complete data set and in the specific subsets (columns). The first row shows the varying distribution of expression values between assembled, but not re-normalized studies (GSEs). The second row shows the variances based on re-normalized expression values. 38

3.5. **Listing of differentially expressed TFs before re-normalization but not differentially expressed afterwards.** Except for Ikarus (italic written), all listed TFs are unknown to have a subtype specific gene expression profile. 41

3.6. **GO enrichment analysis of differentially expressed genes.** Using the enrichment online tool DAVID [155] (online,12/2015). 16 T cell related GO terms where identified as significantly overrepresented in the gene set. Terms are sorted from lowest to highest P-value. 42

3.7. **Enriched GO terms from categories Biological Process (BP) and Molecular Function (MF) for the comparison of differentially expressed genes detected before and after re-normalization.** Enrichment was performed using DAVID Tool [155] (online 12/2015). The set of significantly enriched terms (P-value < 0.05) does not include T cell or lymphocyte specific terms. Terms below the line are not significantly enriched, but enriched 43

4.1. **Overview of feature selection categories and their major properties.** In column four we list exemplary implementations. The table was inspired and adapted from [143, 302]. 47

4.2. **Extracted marker gene candidates** All marker gene candidates identified by feature selection step of analysis pipeline. Bold printed candidate genes were manually selected and evaluated regarding cell type marker functionality, here categorized as surface molecules. Columns "pValue" and "fc(iTreg - nTreg)" contain information from t-test analysis and indicate whether genes are differentially expressed between cell types (double asterix). Table from related publication [198]. 55

4.3. **Performance metrics of validation using SMO classification results for each cell types over all folds.** Metrics were calculated for both candidate sets, six selected marker genes and full set of 41 candidates. Specificity and sensitivity for both classes are "diagonal identical", as a correctly positive inferred instance of one class (*true positive*) is a correctly negative inferred instance (*true negative*) in the other class. . . 57

5.1. **Comparison of steady state data and time-series data** 66

5.2. **Properties of correlation metrics, PCC and SCC [95].** 68

5.3. **Overview of the tools that were discarded from GRN reconstruction comparison.** 11 tools were discarded, because they require time-series data or additional information. Furthermore, no working implementation was found for five tools, while the available implementation of RegnANN is incompatible to required third-party libraries. 76

5.4. **Methods and Tools for GRN inference.** For each method/class of methods (upright written in most left column) we list a number of tools. Underlined tools are used for network inference in Section 5.5.3. Tools marked with an asterisks are not longer supported or not available. . . . 80

5.5. **Number of interactions inferred by the presented tools (rows) tested on ten gene sets (columns).** The table header (bold) shows ten genes sets, the column names indicate the number of genes each set contains. The table is horizontally divided into two parts the upper part lists for each applied tool (rows) the total number of inferred edges per gene set. The lower horizontal part "Evaluation data sets" shows for each evaluation set the number of existing edges based on the given gene set. . 85

5.6. **Performance of GRN reconstruction tools tested using five evaluation set.** The table is vertically divided into four parts, each for one metric, namely sensitivity, specificity, accuracy and F-measure. Tools are shown in the rows of each part, while the columns present the evaluation data set. Best performing tool per metric is highlighted in bold. Numbers marked with an asterics(*) must be interpreted carefully, see text. 87

5.7. **Estimation of potential edges per network.** Here, we compare the maximal possible number of edges per network ("full, directed, without loops") with the number of corresponding edges in evaluation sets. . . . 90

5.8. **Overlap of inferred edges per gene set.** The rows represent the genes sets (10 to 100), while the columns represent the frequency of inference. The column "in total" contains the number of inferred unique edges. For example the column 2× depicts the number of overlapping edges that were inferred by exactly two tools, where 11 indicates how many unique edges were inferred by all applied tools. 90

6.1. **Summary of gene interactions derived from different experiments.** For knock-out (KO) experiments a relation is inferred between the knocked out gene and affected genes. For time-series or steady state experiments interactions are defined between co-regulated or co-expressed genes. (WT stands for wild-type) 98
6.2. **Integrated data sets for Stat6-Th2 cell network reconstruction.** . 100
6.3. **Functions in differentiation processes of STAT6-regulated and STAT6-bound genes encoding for TFs**. The 32 STAT6-regulated TFs were included into a literature search concerning the context of T cell differentiation within the publication period 2000-2015. Table is taken and adapted from [168]. 106

A.1. **Listing of the Treg cell specific microarray experiments included in the meta expression set (see Chapter 3).** Categorization for Treg subtype was performed by biologist for each single microarray. All information were retrieved from NCBI's GEO Platform `http://www.ncbi.nlm.nih.gov/geo`. Full literature references are listed in Table A.2 using PMIDs as key. Table is adapted from [198, in preparation]. 113
A.2. **Literature references of PMIDs corresponding to the experiments used to create the meta-expression set in Chapter 3.** The table is adapted from [198, in preparation]. 118

Acronyms

ANN artificial neural network

APC Antigen-presenting cell

API Application programmer interface

ARACNE Algorithm for the Reconstruction of Accurate Cellular Networks

ASC Anova Sibling Consolidation

BDE Bayesian Dirichlet Equivalence

BIC Bayesian Information Criteria

BN Bayesian network

BoolN Boolean network

CD4 cluster of differentiation 4

CD8 cluster of differentiation 8

CDF chip definition file

CD cluster of differentiation

ChIP-seq Chromatin Immunoprecipitation sequencing

ChIP Chromatin Immunoprecipitation

CLR Context Likelihood of Relatedness

CNV copy number variation

CPDB ConsensusPathDB

DAG directed acyclic graph

DBN Dynamic Bayesian Network

DC dendritic cell

differentially expressed differentially expressed

DIP 5Database of Interacting Proteins

DNA-seq DNA sequencing

DNA deoxyribonucleic acid

DPI Data Processing Inequality

EG Ensembl gene identifier

ENA European Nucleotide Archive

ENSMUSG Ensembl Mouse Gene Identifier

FACS Fluorescence-activated cell sorting

fcs flow cytometry standard

fold change fold change

FNR false negative rate

fpkm fragments per kilobase of exon per million fragments mapped

FPR false positive rate

GDS GEO dataset record

GEO Gene Expression Omnibus

GO Gene Ontology

GPL GEO platform identifier

GRN gene regulatory network

GSEA Gene Set Enrichment Analysis

GSE GEO series record

GSM GEO sample record

GWAS Gene Wide Association Study

HPRD Human Protein Reference Database

Id identifier

IG Information Gain

InsDel insertion and deletion

KEGG Kyoto Encyclopedia of Genes and Genomes

KO knock-out

KI knock-in

LICORN Learning co-operative regulation networks

LIGAP lineage commitment using Gaussian processes

MAS5 Microarray Analysis Suite 5

MHCII Major Histocompatibility Complex class II molecule

MHCI Major Histocompatibility Complex class I molecule

MIAME Minimum Information About a Microarray Experiment

MIM mutual information matrix

miRNA micro RNA

MI mutual information

MM MissMatch

MRMR Minimal Redundancy Maximal Relevance

mRNA messenger RNA

NKC natural killer cell

NCA Network Component Analysis

NGS Next Generation Sequencing

NIR Network Identification by multiple Regression

ODE ordinary differential equation

pBoolN Probabilistic Boolean Network

PCA principle component analysis

PCC Pearson Correlation Coefficient

PMC PubMed Central

PMID PubMed identifier

PM PerfectMatch

PPI protein-protein interaction

RMA Robust Multi-array Average

RNA-seq RNA sequencing

RNA ribonucleic acid

RNN Recurrent Neurall Network

RN Relevance Networks

RT reverse transcriptase

SCC Spearman's rank Correlation Coefficient

SNA social network analysis

SNP single nucleotide polyphormism

SQL structured query language

SRA Sequence Read Archive

SVM Support Vector Machine

TCR T cell receptor

TFBS transcription factor binding site

Tfh follicular T helper

TF transcription factor

Th17 T helper 17

Th1 T helper 1

Th2 T helper 2

Th T helper

TLR Toll-like receptor

TNR true negative rate

TPR true positive rate

Treg regulatory T helper

TSS transcription start side

Weka Weka 3: Data Mining Software

WGCNA Weighted Gene Co-expression Network Analysis

WT wild-type